

# Querying Visible and Invisible Tables in the Presence of Integrity Constraints

Michael Benedikt  
Oxford University, UK

Pierre Bourhis  
CNRS, CRISAL

Balder ten Cate  
UC-Santa Cruz and LogicBlox Inc

Gabriele Puppis  
CNRS, LaBRI

## Abstract

We provide a wide-ranging study of the scenario where a subset of the tables in a relational schema are visible to a user — that is, their complete contents are known — while the remaining tables are invisible. The schema also has a set of integrity constraints, which may relate the visible tables to invisible ones but also may constrain both the visible and invisible instances. We want to determine whether information about a user query can be inferred using only the visible information and the constraints. We consider whether positive information about the query can be inferred, and also whether negative information (the query does not hold) can be inferred. We further consider both the instance-level version of the problem (the visible table extensions are given) and the schema-level version, where we want to know whether information can be leaked in some instance of the schema. Our instance-level results classify the complexity of these problems, both as a function of all inputs, and in the size of the instance alone. Our schema-level results exhibit an unusual dividing line between decidable and undecidable cases.

## 1 Introduction

There are many applications scenarios where a collection of datasources are defined, but a given user or class of users has access to only a subset of these sources. For example, for privacy reasons a data owner may explicitly restrict access to a subset of the stored tables, or to virtual tables defined via queries. Restricted access can also emerge naturally in data integration, where some datasources may be virtual and are defined via mappings to sources. In this case, the virtual tables are not accessible (to the middleware) but the backend sources are. Many of these scenarios can be subsumed by considering a schema consisting of a set of relations related by integrity constraints, with only a subset of the relations accessible. A basic question is whether a given data design of this form renders some information inaccessible. Traditional access control

mechanisms can restrict explicit access, but they can not prevent “information leakage” that may occur due to the presence of semantic relationships either between datasources or within one datasource. For example, if there are referential constraints between relations  $R$  and  $S$  in a database, a designer who wants to restrict users from accessing the information in  $R$  may also have to restrict access to  $S$ .

In this work we consider exactly this scenario, where a set of semantically-related relations are hidden while for another set the complete contents are visible. We will consider semantic relationships specified in a variety of languages that are rich enough to capture complex relationships between sources, including relationships that arise in data integration, as well as common integrity constraints within a single source, such as referential constraints. The basic analysis problem we will consider will be the following: given a schema and a (for simplicity, Boolean) query  $Q$ , can we infer using data and schema information that the result of  $Q$  is true or that the result is false.

**Example 1.** Consider a medical datasource with relation  $\text{Appointment}(p, a, \dots)$  containing patient names  $p$ , appointment ids  $a$ , and other information about the appointment, such as the name of the doctor. A dataowner makes available one projection of  $\text{Appointment}$  by creating a relation  $\text{Patient}(p)$  defined by the constraints:

$$\begin{aligned} \forall p \text{ Patient}(p) &\rightarrow \exists a d \bar{y} \text{ Appointment}(p, a, d, \bar{y}) \\ \forall p a d \bar{y} \text{ Appointment}(p, a, d, \bar{y}) &\rightarrow \text{Patient}(p) . \end{aligned}$$

The query  $Q = \exists a \bar{y} \text{ Appointment}(\text{“Smith”}, a, \text{“Jones”}, \bar{y})$  asking whether patient Smith made an appointment with Dr. Jones will be secure under this schema in one sense: an external user with access to  $\text{Patient}$  will never be sure that the query is true. On the other hand, on an instance where the visible relation  $\text{Patient}$  is empty, an external user will know that the query is false.

We will say that there is a *Negative Query Implication* on the visible instance where  $\text{Patient}$  is empty, since a user can determine whether the query is false.

**Our results.** We will consider the instance-based problems – given a query and instance, can a user determine that the query is either true (a Positive Query Implication) or false (Negative Query Implication). We also look at the corresponding schema-level problem: given a schema, is there *some* instance where a query implication of one of the above types occurs.

We start by observing that the instance-level problems, both positive and negative, are decidable for a very broad class of constraints. However, when we analyze the complexity of the decision problem as the size of the instance increases, we see surprisingly different behavior between the positive and negative case. For very simple constraints, such as inclusion dependencies, the negative query implication problems are very well-behaved as the instance changes, in polynomial time and definable within a well-behaved query language. For the same class of constraints, the corresponding positive query implication questions are hard even when the schema and query are fixed.

When we turn to the schema-level problems, even decidability is not obvious. We prove a set of “critical instance” results, showing that whenever there is an instance where information about the query can be implied, the “obvious instance” works. Thus the schema-level problems reduce to special cases of the instance-level problems. Although we use this technique to obtain decidability and complexity results both for positive and for negative query implication, the classes of constraints to which they apply are different. We give undecidability results that show that when the classes are even slightly enlarged, decidability of the existence of a schema with a query implication is lost.

**Our techniques.** In the process, we introduce a number of tools for use in querying of mixtures of complete and incomplete information.

- **Embeddings in rich decidable logics.** Our first technique involves showing that a large class of instance based problems can be solved by translating them into satisfiability problems within a rich fragment of first-order logic, the guarded negation fragment, and then analyzing recently-developed techniques for analyzing this logic. As we will show, this allows to make use of powerful prior decidability results “off-the-shelf”. But to get tight complexity bounds, we also require a new analysis of the complexity of these logics.
- **Decidability via canonical counterexamples.** The schema-level analysis asks if there is *some* instance on which information about the query can be derived. As mentioned above, we show that whenever there is some instance, it can be taken to be the “simplest possible instance”. While this idea has been used before to simplify analysis of undecidability (e.g. [GM14]), we give a broad result that allows the use of it for decidability.
- **Tractability via Greatest Fixed-point logic.** For our instance-level problems concerning inference of negative information, we introduce a new technique that shows that the problem can be reduced to evaluating a query in *greatest fixedpoint-Datalog* (GFP-Datalog) on the instance. Since GFP-Datalog queries can be evaluated in polynomial time, this shows tractability in the instance size. This is in contrast to methods used in open world query answering based on definability in Datalog, a subset of least fixedpoint logic. The reduction to GFP-Datalog requires a new analysis of when these inference problems are “active-domain controllable” (it suffices to see that the query value is invariant over all hidden databases that lie within the active domain of the visible instance).
- **Relationships between problems.** We prove reductions relating the positive and negative versions of our problems, relating the schema- and instance-level problems, and relating our problems to the widely-studied “certain answer problem”. We apply these reductions to get upper and lower bounds for our problems.

In addition to the techniques above, our lower bound results involve a number of techniques for coding computation in query inference problems.

**Related Work.** Two different communities have studied the problem of de-

terminating the information that can be inferred from complete access to data in a subset of the relations in a relational schema using constraints that relate the subset to the full vocabulary.

In the database community, the focus has been on views. The schema is divided into the “base tables” and “view tables”, with the latter being defined by queries (typically conjunctive queries) in terms of the former. Given a query over the schema, the basic computational problem is determining which answers can be inferred using only the values of the views. Abiteboul and Duschka [AD98] isolate the complexity of this problem in the case where views are defined by conjunctive queries; in their terminology, it is “querying under the Closed World Assumption”, emphasizing the fact that the possible worlds revealed by the views are those where the view tables have exactly their visible content. In our terminology, this corresponds exactly to the “Positive Query Implication” (PQI) problem in the case where the constraints consist entirely of conjunctive query view definitions. Chirkova and Yu [CY14] extend to the case where conjunctive query views are supplemented by weakly acyclic dependencies. Another subcase of PQI that has received considerable attention is the case where the constraints consist *only* of “completeness assertions” between the invisible and visible portions of the schema. A series of papers by Fan and Geerts [FG10a, FG10b] isolate the complexity for several variations of the problem, with particular attention to the case where the completeness assertions are via inclusion dependencies from the invisible to the visible part.

The PQI problem is also related to work on instance-based determinacy (see in particular the results of Howe et al. in [KUB<sup>+</sup>12]) while the “Negative Query Implication” (NQI) problem is studied in the view context by Mendelzon and Zhang [ZM05], under the name of “conditional emptiness”. In both cases, the emphasis has been on view definitions rather than more general constraints which may restrict both the visible and invisible instance. In contrast, in our work we deal with constraint classes that can restrict the visible and invisible data in ways incomparable to view definitions (see also the comparison in Section 5).

In the description logic community, the emphasis has not been on views, but on querying incomplete information with constraints. Our positive query implication problems relate to work in the description logic community on *hybrid closed and open world query answering* or *DBoxes*, in which the schema is divided into closed-world and open-world relations. Given a Boolean CQ, we want to find out if it holds in all instances that can add facts to the open-world relations but do not change the closed-world relations. In the non-Boolean case, the generalization is to consider which tuples from the initial instance are in the query answer on all such instances. Thus closed-world and open-world relations match our notion of visible and invisible, and the hybrid closed and open world query answering problem matches our notion of positive query implication, except that we restrict to the case where the open-world/visible relations of the instance are empty. It is easy to see that this restriction is actually without loss of generality: one can reduce the general case to the case we study with a simple linear time reduction, making a closed-world copy  $R'$  of each open-world

relation  $R$ , and adding an inclusion dependency from  $R'$  to  $R$ . As with the database community, the main distinction between our study of the Positive Query Implication problem and the prior work in the DL community concerns the classes of constraints considered. Lutz et al. [LSW12] study the complexity of this problem for the constraint languages  $\mathcal{EL}$  and DL-LITE, giving a dichotomy between CO-NP-hard and first-order rewritable sets of constraints. They also show that in all the tractable cases, the problem coincides with the classical open-world query answering problem. Franconi et al. [FIS11] show CO-NP-completeness for a disjunction-free description logic. Our results on the data complexity of PQI consider the same problem, but for decidable constraint languages that are more expressive, and in particular, can handle relations of arbitrary arity, rather than arity at most 2 as in [LSW12, FIS11].

In summary, both the database and DL communities considered the Positive Query Implication questions addressed in this paper, but for constraint classes that are different from those we consider. The Negative Query Implication problems are not well-studied in the prior literature, and we know of no work dealing with the schema-level questions (asking for the existence of an instance with a query implication) in prior work. However, in this paper we show (see Subsection 4.2) that there is a close relation between the existence questions to work concerning conservativity and modularity of constraints of Lutz et al. [LW07, KLWW13].

Note that our schema-level analysis considers the existence of *some* instance where the query result can be inferred. In contrast, the work of Miklau and Suciu [MS07] considers whether a “typical” instance allows such inference. We do not deal with probabilistic modelling in this work.

## 2 Definitions

We consider partitioned schemas (or simply, schemas)  $\mathbf{S} = \mathbf{S}_h \cup \mathbf{S}_v$ , where the partition elements  $\mathbf{S}_h$  and  $\mathbf{S}_v$  are finite sets of relation names (or simply, relations), each with an associated arity. These are the *hidden* and *visible* relations, respectively. An *instance* of a schema maps each relation to a set of tuples of the associated arity. Instances will be used as inputs to the computational problems that are the focus of this work – in this case the instances must be finite. Our computational problems also quantify over instances, and they are also well-defined when the quantification is over all (finite or infinite) instances. For simplicity, by default *instances are always finite*. However, as we will show, taking any of the quantification over all instances will never impact our results, and this will allow us to make use of infinite instances freely in our proofs. The *active domain* of an instance is the set of values occurring within the interpretation of some relation in the instance.

As a suggestive notation, we write  $\mathcal{V}$  for instances over  $\mathbf{S}_v$  and  $\mathcal{F}$  for instances over  $\mathbf{S}$ . Given an instance  $\mathcal{F}$  for  $\mathbf{S}$ , its restriction to the  $\mathbf{S}_v$  relations will be referred to as its *visible part*, denoted  $\text{Visible}(\mathcal{F})$ .

We will look at integrity constraints defined by Tuple-generating Dependen-

cies (TGDs), which are first-order logic sentences of the form

$$\forall \bar{x} \phi(\bar{x}) \rightarrow \exists \bar{y} \rho(\bar{x}, \bar{y})$$

where  $\phi$  and  $\rho$  are conjunctions of atoms, which may contain variables and/or constants, and where all the universally quantified variables  $\bar{x}$  appear in  $\phi(\bar{x})$ . For all the problems considered in this work, one can take w.l.o.g. the right-hand side  $\rho$  to consist of a single atom, and we will assume this henceforth. We will often omit the universal quantifiers, writing just  $\phi(\bar{x}) \rightarrow \exists \bar{y} \rho(\bar{x}, \bar{y})$ . Several classes of TGDs will be of particular interest:

- *Linear TGDs*: those where  $\phi$  consists of a single atom.
- *Inclusion Dependencies (IDs)*, linear TGDs where each of  $\phi$  and  $\rho$  have no constants and no repeated variables. These correspond to traditional referential constraints.
- Many of our results on inclusion dependencies will hold for two more general classes. *Frontier-guarded TGDs* (FGTGDs) [BLMS09] are TGDs where one of the conjuncts of  $\phi$  is an atom that includes every universally quantified variable  $x_i$  occurring in  $\rho$ . *Connected TGDs* require only that the *co-occurrence graph* of  $\phi$  is connected. The nodes of this graph are the variables  $\bar{x}$ , and variables are connected by an edge if they co-occur in an atom of  $\phi$ .

Note that every ID is a linear TGD, and every linear TGD is frontier-guarded. We will also consider two constraint languages that are generalizations of FGTGDs.

- We allow disjunction, by considering Disjunctive Frontier-guarded TGDs, which are of the form

$$\forall \bar{x} \phi(\bar{x}) \rightarrow \exists \bar{y} \bigvee_i \rho_i(\bar{x}, \bar{y})$$

where each  $\rho_i$  is a conjunction of atoms and there is one atom conjoined in  $\phi$  that includes every variable  $x_i$  included in some  $\rho_i$ .

- Many of our results apply to an even richer constraint language containing Disjunctive FGTGDs, the *Guarded Negation Fragment*, denoted GNFO. GNFO is built up inductively according to the grammar:

$$\begin{aligned} \phi ::= & R(\bar{t}) \mid t_1 = t_2 \mid \exists x \phi \mid \phi \vee \phi \mid \phi \wedge \phi \mid \\ & R(\bar{t}, \bar{y}) \wedge \neg \phi(\bar{y}) \end{aligned}$$

where  $R$  is either a relation symbol or the equality relation  $x = y$ , and the  $t_i$  represent either variables or constants. Notice that any use of negation must occur conjoined with an atomic relation that contains all the free variables of the negated formula – such an atomic relation is a *guard* of the formula. In database terms, GNFO is equivalent to relational algebra where the *difference operator* can only be used to subtract query results from a relation. The VLDB paper [BtCO12] gives both Relational algebra and SQL-based syntax for GNFO, and argues that it covers useful queries and constraints in practice.

For simplicity (so that all of our constraints are well-defined on instances) we will always assume that our GNFO formulas are domain-independent; to enforce this we can use the relational algebra syntax for capturing these queries, mentioned above. The reader only needs to know a few facts about GNFO. The first is that it is quite expressive, so in proving things about GNFO constraints we immediately get the results for many classes of constraints that we have mentioned above. GNFO contains every positive existential formula, is closed under Boolean combinations of sentences, and it subsumes disjunctive frontier-guarded TGDs up to equivalence. That is, by simply writing out a disjunctive frontier-guarded TGD using  $\exists, \neg, \wedge$ , one sees that these are expressible in GNFO.

Secondly, we will use that GNFO is “tame”, encapsulated in the following result from [BtCS11]:

**Theorem 1** ([BtCS11]). *Satisfiability for GNFO sentences can be tested effectively, and is 2EXPTIME-complete. Furthermore, every satisfiable sentence has a finite satisfying model.*

Note that GNFO does *not* subsume the constraints corresponding to CQ view definitions (e.g.  $A(x, y) \wedge B(y, z) \leftrightarrow V(x, z)$  cannot be expressed in GNFO). However we will cover this special class of constraints in Section 5.

Finally, we will consider *Equality-generating Dependencies* (EGDs), of the form

$$\forall \bar{x} \phi(\bar{x}) \rightarrow x_i = x_j$$

where  $\phi$  is a conjunction of atoms and  $x_i, x_j$  are variables. EGDs generalize well-known relational database constraints, such as functional dependencies and key constraints. *EGDs with constants* further allow equalities between variables and constants, e.g.  $x_i = a$ , in the right-hand side.

For our query language we consider *conjunctive queries* (CQs), first-order formulas built up from relational atoms via conjunction and existential quantification (equivalently, relational algebra queries built via selection, projection, join, and rename operations), and also *unions of CQs* (UCQs), which are disjunctions (relational algebra UNIONS) of CQs. *Boolean UCQs* are simply UCQs with no free variables. Every CQ  $Q$  is associated with a *canonical database*  $\text{CanonDB}(Q)$ , where the domain consists of variables and constants of  $Q$  and the facts are the atoms of  $Q$ .

We will always assume that we have associated with each value a corresponding constant, and we will identify each constant with its value. Thus distinct constants will always be forced to denote distinct domain elements – this is often called the “unique name assumption” (UNA) [AHV95]. While the presence or absence of constants will often make no difference in our results, there are several problems where their presence adds significant complications. In contrast, it is easy to show that the presence of constants without the UNA will never make any difference in any of our results. *Note that in our constraint and query languages above, with the exception of IDs, constants are allowed by default.* When we want to restrict to formulas without constants, we add the

prefix NoConst – e.g. NoConst FGTGD denotes the frontier-guarded TGDs that do not contain constants.

The crucial definition for our work is the following:

**Definition 2.** Let  $Q$  be a Boolean UCQ over schema  $\mathbf{S}$ ,  $\mathcal{C}$  a set of constraints over  $\mathbf{S}$ , and  $\mathcal{V}$  an instance over a visible schema  $\mathbf{S}_v \subseteq \mathbf{S}$ .

- $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$  if for every finite instance  $\mathcal{F}$  satisfying  $\mathcal{C}$ , if  $\mathcal{V} = \text{Visible}(\mathcal{F})$  then  $Q(\mathcal{F}) = \text{true}$ .
- $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$  if for every finite instance  $\mathcal{F}$  satisfying  $\mathcal{C}$ , if  $\mathcal{V} = \text{Visible}(\mathcal{F})$  then  $Q(\mathcal{F}) = \text{false}$ .

We call an  $\mathbf{S}_v$ -instance  $\mathcal{V}$  *realizable* w.r.t.  $\mathcal{C}$  if there is a  $\mathbf{S}$ -instance  $\mathcal{F}$  satisfying  $\mathcal{C}$  such that  $\mathcal{V} = \text{Visible}(\mathcal{F})$ . If an instance  $\mathcal{V}$  is not realizable w.r.t.  $\mathcal{C}$ , then, trivially,  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ . In practice, realizable instances are the only  $\mathbf{S}_v$ -instances we should ever encounter. For simplicity we state our instance-level results for the PQI and NQI problems that take as input an arbitrary instance of  $\mathbf{S}_v$ . But since our lower bound arguments will only involve realizable instances, an alternative definition that assumes realizable inputs yields the same complexity bounds.

$\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  states something about every finite instance, in line with our default assumption that instances are finite. We can also talk about an “unrestricted version” where the quantification is over every (finite or infinite) instance. *For the constraints we deal with, there will be no difference between these notions.* That is, we will show that the finite and unrestricted versions of PQI coincide for a given class of arguments  $Q, \mathcal{C}, \mathbf{S}, \mathcal{V}$ . We express this by saying that “ $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  is finitely controllable”, and similarly for NQI.

Often we will be interested in studying the behavior of these problems when  $Q, \mathbf{S}$  and  $\mathcal{C}$  are fixed, e.g. looking at the computation time varies in the size of  $\mathcal{V}$  only. We refer to this as the *data complexity* of the PQI, (resp. NQI) problem.

The PQI problem contrasts with the usual *Open-World Query Answering* or *Certain Answer* problem, denoted here  $\text{OWQ}(Q, \mathcal{C}, \mathcal{F})$ , which is studied extensively in databases and description logics. The latter problem takes as input a Boolean query  $Q$ , an instance  $I$ , and a set of constraints  $\mathcal{C}$ , and returns **true** iff the query holds in any finite instance  $I'$  containing all facts of  $I$ . In PQI (and NQI) we further constrain the instance to be fixed on the visible part while requiring the invisible part of the input instance to be empty. This is the mix of “Closed World” and “Open World”, and we will see that this Closed World restriction can make the complexity significantly higher.

**Example 2.** Consider a schema with inclusion dependencies  $F_1(x) \rightarrow \exists y U(x, y)$  and  $U(x, y) \rightarrow F_2(y)$ , where  $F_1$  and  $F_2$  are visible but  $U$  is not. Consider the query  $Q = \exists x U(x, x)$  and instance consisting only of facts  $F_1(a), F_2(a)$ .

There is a PQI on this instance, since  $F_1(a)$  implies that  $U(a, c)$  holds for some  $c$ , but the other constraint and the fact that  $F_2$  must hold only of  $a$  means that  $c = a$ , and hence  $Q$  holds.

In contrast, one can easily see that  $Q$  is not certain in the usual sense, where  $F_1$  and  $F_2$  can be freely extended with additional facts.



Our schema-level problems concern determining if there is a realizable instance that admits a query implication:

**Definition 3.** For  $Q$  a Boolean conjunctive query over schema  $\mathbf{S}$ , and  $\mathcal{C}$  a set of constraints over  $\mathbf{S}$ , we let:

- $\exists \text{PQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  if there is a realizable  $\mathbf{S}_v$ -instance  $\mathcal{V}$  such that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ ;
- $\exists \text{NQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  if there is a realizable  $\mathbf{S}_v$ -instance  $\mathcal{V}$  such that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ .

Note that these problems now quantify over instances twice, and hence there are alternatives depending on whether the instance  $\mathcal{V}$  is restricted to be finite, and whether the hidden instances  $\mathcal{F}$  are restricted to be finite. For a class of input  $Q, \mathcal{C}, \mathbf{S}$ , we say that  $\exists \text{PQI}(Q, \mathcal{C}, \mathbf{S})$  is “finitely controllable” if in both quantifications, quantification over finite instances can be freely replaced with quantification over arbitrary instances.

### 3 Positive Query Implication

#### 3.1 Instance problems

We begin with a study of PQI. We will show that PQI is decidable for the rich constraint language GNFO, the guarded negation fragment, which includes guarded TGDs, disjunctive guarded TGDs, and Boolean combinations with Boolean CQs. The key is that we can translate the PQI problem to a satisfiability problem for GNFO.

**Theorem 4.** The problem  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$ , as  $Q$  ranges over Boolean UCQs and  $\mathcal{C}$  over GNFO constraints, is in 2EXPTIME.

Furthermore, for such constraints the problem is finitely controllable, that is,  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$  iff for every instance  $\mathcal{F}$  (of any size) satisfying  $\mathcal{C}$ , if  $\mathcal{V} = \text{Visible}(\mathcal{F})$ , then  $Q(\mathcal{F}) = \text{true}$ .

*Proof.* We just note that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  translates to unsatisfiability of the following formula:

$$\phi_{Q, \mathcal{C}, \mathbf{S}, \mathcal{V}}^{\text{PQItoGNF}} = \neg Q \wedge \mathcal{C} \wedge \bigwedge_{R \in \mathbf{S}_v} \left( \bigwedge_{R(\bar{a}) \in \mathcal{V}} R(\bar{a}) \wedge \forall \bar{x} \left( R(\bar{x}) \rightarrow \bigvee_{R(\bar{a}) \in \mathcal{V}} \bar{x} = \bar{a} \right) \right)$$

If the constraints are in GNFO, then the formula above is also in GNFO. The finite controllability of  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  comes from the finite controllability of GNFO formulas (Theorem 1)  $\square$

Above we are using results on satisfiability of GNFO as a “black-box”. Satisfiability tests for GNFO work by translating a satisfiability problem for a formula into a tree automaton which must be tested for non-emptiness. By a finer analysis of this translation of GNFO formulas to automata, we can see that the *data complexity* of the problem is only singly-exponential.

**Theorem 5.** *If  $Q$  is a Boolean UCQ and  $\mathcal{C}$  is a conjunction of GNFO constraints over a schema  $\mathbf{S}$ , then the data complexity of  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  (that is, as  $\mathcal{V}$  varies over instances) is in EXPTIME.*

*Proof.* We sketch a proof of this, which involves a finer analysis of the conversion of GNFO formulas to automata. We define a variant of the normal form introduced in [BtCS11], called *GN-normal form*, via the following grammar:

$$\begin{aligned}\varphi &::= \bigvee_i \exists \bar{x} \left( \bigwedge_j \psi_{ij} \right) \\ \psi &::= \alpha(\bar{x}) \mid \alpha(\bar{x}) \wedge \varphi(\bar{x}) \mid \alpha(\bar{x}) \wedge \neg \varphi(\bar{x})\end{aligned}$$

where  $\alpha(\bar{x})$  is an atomic formula. Further, let us say that a GNFO formula  $\phi$  is *equality-normalized* if

- (i) every occurrence of  $R(\bar{t})$  in  $\phi$  appears in conjunction with

$$\text{distinct}(\bar{t}) := \bigwedge_{t \in \bar{t}} (t = t) \wedge \bigwedge_{\substack{t, t' \in \bar{t} \\ t \neq t'}} \neg(t = t'),$$

- (ii) whenever equalities are used as guards for negations, then these equality guards are of the form  $x = x$ ,
- (iii) every occurrence of equality in  $\phi$  is either an equality comparison of constants, or comes from (i) or (ii).

The *width* of  $\phi$ , denoted  $\text{width}(\phi)$ , is the maximum number of free variables of any subformula of  $\phi$ .

Let  $\psi$  be a GNFO sentence, with  $s = |\psi|$ . We can construct a (DAG representation of an) equi-satisfiable equality-normalized GN-normal form sentence  $\varphi$  such that:

- the size of  $\varphi$  is at most  $2^{f(s)}$ ,
- $\text{width}(\varphi) \leq s$

where  $f$  is a polynomial function independent of  $\psi$ . For  $\phi'$  in GN-normal form, we define  $\text{rank}_{\text{CQ}}(\phi')$  to be the maximum number of conjuncts  $\psi_i$  in any CQ-shaped subformula  $\exists \bar{x} \bigwedge_i \psi_i$  of  $\phi'$ , for non-empty  $\bar{x}$ .

The construction is not difficult, and a more general statement can be found in Proposition 31 of the pre-print available at [BCtCB15].

The following key proposition, shows that formulas in GN-normal form can be translated into automata with size controlled by the CQ-rank and width:

**Proposition 6.** *For every GNFO formula  $\phi'$  in GN-normal form, there is an alternating two-way parity automaton  $T_{\phi'}$  on infinite trees such that  $T_{\phi'}$  recognizes a non-empty language iff  $\phi'$  is satisfiable. Moreover, the number of states of the automaton, and the running time needed to form it, is bounded by  $f(s') \cdot 2^{f(w'r')}$ , where  $s' = |\phi'|$ ,  $w' = \text{width}(\phi')$ ,  $r' = \text{rank}_{\text{CQ}}(\phi')$ , and  $f$  is a polynomial function independent of  $\phi'$ .*

Proposition 6 is proven by creating an automaton whose state set consists of a collection of formulas derived from  $\phi'$ . If the formula was guarded, these would just be subformulas, but for CQ-shaped subformulas, one will have to throw in all subformulas, representing guesses as which of the conjuncts were true at a given bag of a tree-like structure. The details are again in the preprint at [BCtCB15], see Corollary 10 (formation of an automaton based on the closure) and Lemma 24 (closure size as a function of CQ-rank).

Now fix a Boolean UCQ  $Q$  and a conjunction  $\mathcal{C}$  of GNFO constraints over a schema  $\mathbf{S}$ . Without loss of generality, we can assume that the constraints in  $\mathcal{C}$  are already in GN-normal form. Consider the formula  $\phi_{Q,\mathcal{C},\mathbf{S},\mathcal{V}}^{\text{PSBtoGNF}}$  in the proof of Theorem 4:

$$\neg Q \wedge \mathcal{C} \wedge \bigwedge_{R \in \mathbf{S}_v} \left( \bigwedge_{R(\bar{a}) \in \mathcal{V}} R(\bar{a}) \wedge \forall \bar{x} \left( R(\bar{x}) \rightarrow \bigvee_{R(\bar{a}) \in \mathcal{V}} \bar{x} = \bar{a} \right) \right).$$

Note that this formula satisfies all the conditions of the GN-normal form but those related to equality. Nonetheless, we can further break up the ‘dangerous’ subformulas of the form  $\bigvee_{R(\bar{a}) \in \mathcal{V}} \bar{x} = \bar{a}$ , grouping based on the repetition pattern in  $\bar{a}$  and adding inequalities between  $\bar{x}$  that match the non-repeated positions in each group. We can also conjoin each equality with a guard  $R(\bar{x})$ . With this linear-time transformation, the conditions for normalizing equalities will be satisfied.

Thus the formula  $\phi_{Q,\mathcal{C},\mathbf{S},\mathcal{V}}^{\text{PSBtoGNF}}$  can be normalized in polynomial time, and the width and  $\text{rank}_{\text{CQ}}$  of  $\phi_{Q,\mathcal{C},\mathbf{S},\mathcal{V}}^{\text{PSBtoGNF}}$  are fixed when  $Q$ ,  $\mathcal{C}$ , and  $\mathbf{S}$  are fixed. Applying Proposition 6, we get a polynomial-sized two-way alternating automaton. Since emptiness of such automata can be checked in EXPTIME [Var98], the bound claimed in the theorem now follows.  $\square$

The data complexity bound in Theorem 5 is tight even for inclusion dependencies. The proof, proceeds by showing that a “universal machine” for alternating PSPACE can be constructed by fixing appropriate  $Q, \mathcal{C}, \mathbf{S}$  in a PQI problem.

**Theorem 7.** *There are a Boolean UCQ  $Q$  and a set  $\mathcal{C}$  of IDs over a schema  $\mathbf{S}$  for which the problem  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  is EXPTIME-hard in data complexity.*

*Proof.* We first prove the hardness result using a UCQ  $Q$ ; later, we show how to generalize this to a CQ. We reduce the acceptance problem for an alternating PSPACE Turing machine  $M$  to the negation of  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$ .

A configuration of  $M$  is defined, as usual, by a control state, a position of the head on the tape, and a finite string representing the content on the tape, which is assumed to be empty at the beginning. We distinguish between existential and universal control states of  $M$ . The transition function of  $M$  describes a set of target configurations on the basis of the current configuration and, without loss of generality, we assume that every set of target configurations has cardinality 2. The computation of  $M$  is represented by a tree of configurations, where the root represents the initial configuration and where every configuration with

an existential control state (resp., a universal control state) has exactly one successor configuration (resp., two successor configurations). To make the coding simpler we need to adopt a non-standard acceptance condition. Specifically, we assume that the Turing machine  $M$  never halts, namely, its transition function is defined on every configuration, and we distinguish two special control states,  $q_{\text{acc}}$  and  $q_{\text{rej}}$ . When the machine reaches one of these two states in a configuration, it loops forever without changing the configuration. We say that  $M$  accepts (the empty input) if for all paths in the computation tree, the state  $q_{\text{acc}}$  is eventually reached; symmetrically, we say that  $M$  rejects if there is at least one path in the computation tree that contains the state  $q_{\text{rej}}$ . Furthermore, we assume that the  $M$  begin its computation with the head in the second cell and never visits the extremal positions of its tape (this can be easily enforced by marking the second position of the tape and requiring that whenever the marked position is visited, the machine moves to the right). This latter assumption will simplify checking that two subsequent configurations are correct with respect to the transition rules of  $M$ .

The general idea of the reduction is to create a schema, constraints, and query that together represent a “universal machine” for alternating PSPACE. Given an alternating PSPACE machine encoded in the visible instance, an accepting run is “computed” as an arbitrary full instance satisfying the constraints and violating the query — that is, a witness of the failure of PQI.

We first devise a schema that includes hidden relations that will store the computation tree of a generic alternating PSPACE machine. The constraints and the query will be used to restrict the hidden relations so as to guarantee that the encoding of the computation tree is correct. By “generic” we mean that the hidden relations and corresponding constraints will be independent of the tape size, number of control states, and transition function of the machine. The visible instance will store the “representation” of an alternating PSPACE machine  $M$  — that is, an encoding of  $M$  that can be calculated efficiently once  $M$  is known. This will include the tape size and an encoding of the transition function.

We will then give the reduction that takes an alternating polynomial space machine  $M$  and instantiates all the visible relations with the encoding. The space bound on  $M$  will allow us to create the tape size components of the visible instance efficiently. In contrast, the hidden relations will store aspects of a computation that can *not* be computed easily from  $M$ .

In summary, below we will describe each part of the schema **S** for computation trees of a machine, along with the polynomial mapping that transforms a machine  $M$  into data filling up the visible parts of the schema.

First, we encode the tape (devoid of its content) into a binary relation  $T$ .  $T$  will be visible, and can be filled efficiently once an input  $M$  is known. Given  $M$ , it will be filled in the following natural way: it contains all the facts  $T(y, y')$ , where  $y$  is the identifier of a cell and  $y'$  is the identifier of the right-successor of this cell in the tape. Recall that the input machines  $M$  works on a tape of polynomial length, and hence the visible instance for the relation  $T$  has also

size polynomial in  $M$ . Despite the fact that the tape length is finite, we need every cell to have a successor; this will be exploited later to detect badly formed encodings of configurations. We thus include the fact  $T(y, y)$  in the instance of the visible relation  $T$ , where  $y$  is the identifier of the rightmost cell of the tape. For similar reasons, we add another visible relation  $T_0$ , intended to distinguish the first two cells in the tape. We will form  $T_0$  from  $M$  by filling it with the identifiers of the first two cells in the tape.

As for the configurations of the machine, these are described by specifying, for each configuration and each tape cell, a suitable value that describes the content of that cell, together with the information on whether the Turing machine has its head on the cell, and what is the corresponding control state. For a technical reason (specifically, to allow detecting violations of the transition rules between pairs of subsequent configurations), we adjoin to the labelling of a cell also that of the adjacent cells whenever the head is within the neighbourhood. Formally, the configurations of the machine are encoded by a hidden ternary relation  $C$ , where each fact  $C(x, y, z)$  indicates that, in the configuration identified by  $x$ , the cell  $y$  has value  $z$ .

We enforce the fact that the cell values range over an appropriate domain by a visible unary relation  $V$ . As with all of our visible relations, we can fill  $V$  easily once we have a specific input machine  $M$ . In our reduction from machine  $M$ , we will fill this relation  $V$  with  $(\Sigma \times \Sigma \times Q \times \Sigma) \uplus (\Sigma \times Q \times \Sigma) \uplus (\Sigma \times \Sigma \times Q) \uplus \Sigma$ , where  $\Sigma$  is the tape alphabet of  $M$  and  $Q$  is the set of its control states. If a cell has value  $(a, b, q, c)$ , this means that its content is  $b$ , the Turing machine stores the control state  $q$ , has the head precisely on this cell, and the neighbouring cells to the left and to the right have labels  $a$  and  $c$ , respectively. Similarly, if a cell has value  $(a, q, b)$  (resp.,  $(b, c, q)$ ), this means that its content is  $b$ , the Turing machine stores the control state  $q$ , and the head is on the left-successor (resp., right-successor), which carries the letter  $a$  (resp.,  $c$ ). In all other cases, we simply store the content  $b$  of the cell.

Recall that the tape of the Turing machine will be encoded in the visible relations  $T$  and  $T_0$ . Because we need to associate the same tape with several different configurations, the content of  $T$  and  $T_0$  will end up being replicated within new hidden ternary relations  $T^C$  and  $T_0^C$ , where it will be paired with the identifier of a configuration. Intuitively, a fact  $T^C(x, y, y')$  will indicate that, in the configuration  $x$ , the cell  $y$  precedes the cell  $y'$ . Similarly,  $T_0^C(x, y, y')$  will indicate that the first two cells of the configuration  $x$  are  $y$  and  $y'$ . Of course, we will enforce the condition that the relations  $T^C$  and  $T_0^C$ , projected onto the last two attributes, are contained in  $T$ .

We now turn to the encoding of the computation tree. For this, we introduce a visible unary relation  $C_0$  that will contain the identifier of the initial configuration. We also introduce the hidden binary relations  $S^\exists$ ,  $S_1^\forall$ , and  $S_2^\forall$ . We recall that every configuration  $x$  with an existential control state has exactly one successor  $x'$  in the computation tree, so we represent this with the fact  $S^\exists(x, x')$ . Symmetrically, every configuration  $x$  with a universal control state has exactly two successors  $x_1$  and  $x_2$  in the computation tree, and we represent this with the facts  $S_1^\forall(x, x_1)$  and  $S_2^\forall(x, x_2)$ .

So far, we have introduced the visible relations  $T, T_0, V, C_0$  and the hidden relations  $C, T^C, T_0^C, S^\exists, S_1^\forall, S_2^\forall$ . These are sufficient to store an encoding of the computation tree of the machine. However, the constraint language only allows inclusion dependencies, which are not powerful enough to guarantee that these relations indeed represent a correct encoding. To overcome this problem, we will later introduce a few additional relations and exploit a union of CQs to detect the possible violations of the constraints.

We now list some inclusion dependencies in  $\mathcal{C}$  that enforce basic constraints on the relations.

- We begin with the constraints on the ordering of the cells in the tape:

$$\begin{aligned} T_0(y, y') &\rightarrow T(y, y') & T_0^C(x, y, y') &\rightarrow T^C(x, y, y') \\ T^C(x, y, y') &\rightarrow T(y, y') & T^C(x, y, y') &\rightarrow \exists y'' T^C(x, y', y'') . \\ T_0^C(x, y, y') &\rightarrow T_0(y, y') \end{aligned}$$

- We proceed by enforcing the constraints on the cell values:

$$T^C(x, y, y') \rightarrow \exists z C(x, y, z) \quad C(x, y, z) \rightarrow V(z) .$$

- We finally enforce a tree structure on the configurations assuming that the machines starts with an existential state:

$$\begin{aligned} C_0(x) &\rightarrow \exists x' S^\exists(x, x') & S^\exists(x, x') &\rightarrow \exists x_1 S_1^\forall(x', x_1) \\ S^\exists(x, x') &\rightarrow \exists y y' T_0^C(x, y, y') & S^\exists(x, x') &\rightarrow \exists x_2 S_2^\forall(x', x_2) \\ S_1^\forall(x, x_1) &\rightarrow \exists y y' T_0^C(x, y, y') & S_1^\forall(x, x_1) &\rightarrow \exists x' S^\exists(x_1, x') \\ S_2^\forall(x, x_2) &\rightarrow \exists y y' T_0^C(x, y, y') & S_2^\forall(x, x_2) &\rightarrow \exists x' S^\exists(x_2, x') . \end{aligned}$$

Next, we explain how to detect badly-formed encodings of the computation tree. For this, we use additional visible binary relations  $\text{Err}_{\perp,0}, \text{Err}_{\perp}, \text{Err}_C, \text{Err}_{S^\exists}, \text{Err}_{S_1^\forall},$  and  $\text{Err}_{S_2^\forall}$ , instantiated as follows.

- The relation  $\text{Err}_{\perp,0}$  contains all the pairs in  $V \times V$ , but the pair  $(z_0, z_1)$ , where  $z_0$  is the cell value  $(\perp, \perp, q_0)$ ,  $z_1$  is the cell value  $(\perp, \perp, q_0, \perp)$ ,  $q_0$  is the initial state of the Turing machine  $M$  and  $\perp$  is the blank tape symbol. Intuitively,  $\text{Err}_{\perp,0}$  contains precisely those pairs of values that cannot be associated with the first two cells in the initial configuration (recall that  $M$  starts with the head on the second cell). Similarly, the relation  $\text{Err}_{\perp}$  contains all the pairs in  $V \times V$  but the following ones: the previous pair  $(z_0, z_1)$ , the pair  $(z_1, z_2)$ , where  $z_2 = (\perp, q_0, \perp)$ , the pair  $(z_2, z_3)$ , where  $z_3 = \perp$ , and the pair  $(z_3, z_3)$ . Namely, the pairs in  $\text{Err}_{\perp}$  are precisely those that cannot be associated with any two adjacent cells in the initial configuration. Accordingly, we can detect whether the initial configuration

is badly-formed using a disjunction of the following CQs:

$$Q_{\perp,0} = \exists x y y' z z' C_0(x) \wedge T_0^C(x, y, y') \wedge C(x, y, z) \wedge C(x, y', z') \wedge \text{Err}_{\perp,0}(z, z')$$

$$Q_{\perp} = \exists x y z u v C_0(x) \wedge T^C(x, y, y') \wedge C(x, y, z) \wedge C(x, y', z') \wedge \text{Err}_{\perp}(z, z') .$$

- The relation  $\text{Err}_C$  contains all pairs of cell values from  $V \times V$  that cannot be adjacent, in any configuration (for example, it contains the pair  $(z, z')$ , where  $z = (a, b, q, c)$  and  $z' = (b', q, c)$ , with  $b \neq b'$ ). Accordingly, a violation of the adjacency constraint on two consecutive cells of some configuration can be detected by the following CQ:

$$Q_C = \exists x y y' z z' T^C(x, y, y') \wedge C(x, y, z) \wedge C(x, y', z') \wedge \text{Err}_C(z, z') .$$

- The relation  $\text{Err}_{S^3}$  contains all pairs of cell values from  $V \times V$  that cannot appear on the same position of the tape at an existential configuration and its immediate successor (this relation is constructed using the transition function of the Turing machine). A violation of the corresponding constraint can be exposed by the following CQ:

$$Q_{S^3} = \exists x x' y z z' S^3(x, x') \wedge C(x, y, z) \wedge C(x', y, z') \wedge \text{Err}_{S^3}(z, z') .$$

- Similarly, the relation  $\text{Err}_{S_1^\forall}$  (resp.,  $\text{Err}_{S_2^\forall}$ ) contains those pairs of values  $(z, z')$  that cannot appear on the same position of the tape of a universal configuration and that of the first (resp., second) successor. The corresponding CQs are

$$Q_{S_1^\forall} = \exists x x_1 y z z' S_1^\forall(x, x_1) \wedge C(x, y, z) \wedge C(x_1, y, z') \wedge \text{Err}_{S_1^\forall}(z, z')$$

$$Q_{S_2^\forall} = \exists x x_2 y z z' S_2^\forall(x, x_2) \wedge C(x, y, z) \wedge C(x_2, y, z') \wedge \text{Err}_{S_2^\forall}(z, z') .$$

It now remains to check that the Turing machine  $M$  reaches the rejecting state  $q_{\text{rej}}$  along some path of its computation tree. This can be done by introducing a last visible relation  $V_{\text{rej}}$  that contains all cell values of the form  $(a, b, q_{\text{rej}}, c)$ , for some  $a, b, c \in \Sigma$ . The CQ that checks this property is

$$Q_{\text{rej}} = \exists x y z C(x, y, z) \wedge V_{\text{rej}}(z) .$$

The final query is thus a disjunction of all the above CQs:  $Q = Q_{\perp,0} \vee Q_{\perp} \vee Q_C \vee Q_{S^3} \vee Q_{S_1^\forall} \vee Q_{S_2^\forall} \vee Q_{\text{rej}}$ .

We are now ready to give the reduction. Denote by  $\mathcal{V}_M$  the instance that captures the intended semantics of the visible relations  $T, T_0, V, C_0, \text{Err}_{\perp,0}, \text{Err}_{\perp}, \text{Err}_C, \text{Err}_{S^3}, \text{Err}_{S_1^\forall}$ , and  $\text{Err}_{S_2^\forall}$ . We have described these semantics above, and argued why they can be created in polynomial time.

Below, we prove that the Turing machine  $M$  has a successful computation tree (where all paths visit the control state  $q_{\text{acc}}$ ) iff  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}_M) = \text{false}$ .

Suppose that  $M$  has a successful computation tree  $\rho$ . On the basis of  $\rho$ , and by following the intended semantics of the hidden relations  $C, T^C, T_0^C, S^\exists, S_1^\forall, S_2^\forall$ , we can easily construct a full instance  $\mathcal{F}$  that satisfies the constraints in  $\mathcal{C}$ , and which agrees with  $\mathcal{V}_M$  on the visible part. Furthermore, because we correctly encode a successful computation tree of  $M$ , the instance  $\mathcal{F}$  violates every CQ of  $Q$ , and hence  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}_M) = \text{false}$ .

Conversely, suppose that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}_M) = \text{false}$ . By Proposition 12, we know  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_M)$  contains an  $\mathbf{S}$ -instance  $\mathcal{F}$  that violates the UCQ  $Q$ . By construction, this instance  $\mathcal{F}$  satisfies the constraints in  $\mathcal{C}$  and agrees with  $\mathcal{V}_M$  on the visible part. We show that the instance  $\mathcal{F}$  witnesses the fact that  $M$  has a successful run. In doing so, we can exploit the fact that  $\mathcal{F}$  is constructed using the chase procedure; in particular, the hidden relations  $S^\exists, S_1^\forall, S_2^\forall$  have a tree-shaped structure, in which every configuration is represented by a unique identifier. The identifier  $x_0$  of the initial configuration is explicitly given in the visible relation  $C_0$ . The content of the first cell of this initial configuration can be easily derived from the series of inclusion dependencies  $C_0(x) \rightarrow \exists x' S^\exists(x, x'), S^\exists(x, x') \rightarrow \exists y, y' T_0^C(x, y, y'), T_0^C(x, y, y') \rightarrow T^C(x, y, y'), T^C(x, y, y') \rightarrow \exists z C(x, y, z)$ . Note that the fact that the CQs  $Q_{\perp,0}$  and  $Q_\perp$  are violated, guarantees that the content of this initial configuration is as expected. Similarly, one can derive the content of the remaining cells by inductively applying the constraints  $T^C(x, y, y') \rightarrow \exists y'' T^C(x, y', y'')$  and  $T^C(x, y, y') \rightarrow \exists z C(x, y, z)$ , and by recalling that the CQ  $Q_C$  is violated. As for the successor configuration(s), one can discover their identifier(s) using the constraints with  $S^\exists, S_1^\forall, S_2^\forall$  in the right-hand side, and applying similar arguments as before. The fact that the CQs  $Q_{S^\exists}, Q_{S_1^\forall}, Q_{S_2^\forall}$  are violated guarantees that the resulting structure of configurations is a correct computation tree of  $M$ . Finally, because the CQ  $Q_{\text{rej}}$  is also violated, the computation tree of  $M$  must be successful.

We have just shown the EXPTIME hardness result for the data complexity of the PQI problem, using a UCQ as query. To finish the proof of Theorem 7, we show that PQI problems for UCQs can be reduced to the analogous problems for CQs.

**Lemma 8.** *Let  $Q = \bigcup Q_i$  be a Boolean UCQ, let  $\mathcal{C}$  be a set of constraints over a schema  $\mathbf{S}$ , and let  $\mathcal{V}$  be an instance for the visible part of  $\mathbf{S}$ . There exist a schema  $\mathbf{S}'$ , a CQ  $Q'$ , a set  $\mathcal{C}'$  of constraints, and an  $\mathbf{S}'_v$ -instance  $\mathcal{V}'$ , all having linear size with respect to the original objects  $\mathbf{S}, Q, \mathcal{C}$ , and  $\mathcal{V}$ , such that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$  iff  $\text{PQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}') = \text{true}$ .*

*Moreover, the transformation preserves all constraint languages considered in our results (e.g., inclusion dependencies).*

*Proof.* The general idea is as follows. For every visible (resp., hidden) relation  $R$  of  $\mathbf{S}$  of arity  $k$ , we add to  $\mathbf{S}'$  a corresponding visible (resp., hidden) relation  $R'$  of arity  $k + 1$ . The idea is that the additional attribute of  $R'$  represents a



truth value, e.g. 0 or 1, which indicates the presence of a tuple in the original relation  $R$ . For example, the fact  $R'(\bar{a}, 1)$  indicates the presence of the tuple  $\bar{a}$  in the relation  $R$ . The constraints and the conjunctive queries in  $Q$  will be rewritten accordingly, so as to propagate these truth values. Thanks to this, we can simulate the disjunctions in the query  $Q$  by using conjunctions and an appropriate look-up table  $\text{Or}$ .

Formally, the new schema  $\mathbf{S}'$  contains a copy  $R'$  of each relation  $R$  in  $\mathbf{S}$ , where  $R'$  is visible iff  $R$  is visible, plus the visible relations  $\text{Or}$ ,  $\text{Zero}$ , and  $\text{One}$  of arities 3, 0, and 0, respectively. For each constraint

$$R(\bar{x}) \rightarrow \exists \bar{y} S_1(\bar{z}_1) \wedge \dots \wedge S_m(\bar{z}_m)$$

in  $\mathcal{C}$ , where  $\bar{z}_1, \dots, \bar{z}_m$  are sequences of variables or constants from  $\bar{x}, \bar{y}$ , we add to  $\mathcal{C}'$  a corresponding constraint of the form

$$R'(\bar{x}, b) \rightarrow \exists \bar{y} S'_1(\bar{z}_1, b) \wedge \dots \wedge S'_m(\bar{z}_m, b) .$$

Similarly, every Boolean CQ  $Q_i = \exists \bar{y} S_1(\bar{z}_1) \wedge \dots \wedge S_m(\bar{z}_m)$  of  $Q$  is rewritten as  $Q'_i(b) = \exists \bar{y} S'_1(\bar{z}_1, b) \wedge \dots \wedge S'_m(\bar{z}_m, b)$ . Let  $n$  be the number of CQs in  $Q$ . We define the Boolean CQ

$$Q' = \exists b_1 \dots b_n, c_0 \ c_1 \dots c_n \ \text{Zero}(c_0) \wedge \bigwedge_i (Q'_i(b_i) \wedge \text{Or}(c_{i-1}, b_i, c_i)) \wedge \text{One}(c_n) .$$

Finally, we construct the visible instance  $\mathcal{V}'$  as follows. We choose some fresh values 0, 1, and  $\perp$  that do not belong to the active domain of  $\mathcal{V}$ . First, we include in  $\mathcal{V}'$  the facts  $\text{Or}(1, 1, 1)$ ,  $\text{Or}(1, 0, 1)$ ,  $\text{Or}(0, 1, 1)$ ,  $\text{Zero}(0)$ , and  $\text{One}(1)$ . Then, for each visible relation  $R$  of  $\mathbf{S}$ , we add to  $\mathcal{V}'$  the fact  $R'(\perp, \dots, \perp, 0)$ , as well as every fact of the form  $R(\bar{a}, 1)$ , where  $R(\bar{a})$  is a fact in  $\mathcal{V}$ . Note that the presence of the facts  $R'(\perp, \dots, \perp, 0)$  in  $\mathcal{V}'$  guarantees that the rewritten CQs  $Q'_i(b)$  can always be satisfied by letting  $b = 0$  and by extending  $\mathcal{V}'$  with hidden facts of the form  $S'(\perp, \dots, \perp, 0)$ .

We are now ready to prove that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$  iff  $\text{PQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}') = \text{true}$ . Suppose that  $\text{PQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}') = \text{true}$  and consider an  $\mathbf{S}$ -instance  $\mathcal{F}$  that satisfies the constraints in  $\mathcal{C}$  and such that  $\text{Visible}(\mathcal{F}) = \mathcal{V}$ . Without loss of generality, we can assume that the active domain of  $\mathcal{F}$  does not contain the values 0, 1, and  $\perp$ . We can easily transform  $\mathcal{F}$  into an  $\mathbf{S}'$ -instance  $\mathcal{F}'$ , by simply expanding all facts with an additional attributed valued 1 and by adding new facts of the form  $R'(\perp, \dots, \perp, 0)$ , for all relations  $R' \in \mathbf{S}'$ , and new visible facts  $\text{Or}(1, 1, 1)$ ,  $\text{Or}(1, 0, 1)$ ,  $\text{Or}(0, 1, 1)$ ,  $\text{Zero}(0)$ , and  $\text{One}(1)$ . Note that  $\text{Visible}(\mathcal{F}') = \mathcal{V}'$ . Since  $\text{PQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}') = \text{true}$ , we know that  $\mathcal{F}'$  satisfies the query  $Q'$ . In particular, it must satisfy a CQ  $Q'_i(b)$  when  $b = 1$ , and this implies that  $\mathcal{F}$  satisfies the UCQ  $Q = \bigvee_i Q_i$ . Conversely, suppose that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$  and consider an  $\mathbf{S}'$ -instance  $\mathcal{F}'$  that satisfies the constraints in  $\mathcal{C}'$  and such that  $\text{Visible}(\mathcal{F}') = \mathcal{V}'$ . By selecting from  $\mathcal{F}'$  only the facts of the form  $R'(\bar{a}, 1)$ , with  $R \in \mathbf{S}$ , and and by projecting away the last attribute, we obtain an  $\mathbf{S}$ -instance  $\mathcal{F}$  that satisfies the constraints in  $\mathcal{C}$  and such that  $\text{Visible}(\mathcal{F}) = \mathcal{V}$ . Finally, since  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ , we know that  $\mathcal{F}$  satisfies at least one CQ  $Q_i$  of  $Q$ , and hence  $\mathcal{F}'$  satisfies  $Q'$ .  $\square$

Applying the lemma above, we have proven Theorem 7.  $\square$

We note that this data complexity lower bound requires a schema with arity above 2. It thus contrasts with results of Franconi et al. [FIS11] that show that the data complexity lies in CO-NP for arity 2. In fact, if we move up from IDs to linear TGDs, we can adapt the argument for the above result to show EXPTIME-hardness even for arity 2.

We show that the 2EXPTIME combined complexity upper bound is tight even for IDs.

**Theorem 9.** *Checking  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$ , where  $Q$  ranges over CQs and  $\mathcal{C}$  over sets of inclusion dependencies, is 2EXPTIME-hard for combined complexity.*

*Proof.* This proof builds up on ideas of the previous proof for Theorem 7. Specifically, we reduce the acceptance problem for an alternating EXPSPACE Turing machine  $M$  to the negation of  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$ , where  $Q$  is a Boolean UCQ and  $\mathcal{C}$  consists of inclusion dependencies. Note that to further reduce the problem to a Positive Query Implication problem with a Boolean CQ, one can exploit Lemma 8.

The additional technical difficulty here is to encode a tape of exponential size. Of course, this cannot be done succinctly using an instance with visible relations. We can however represent the exponential tape by the leaves of a full binary tree. More precisely, we fix an alternating EXPSPACE Turing machine  $M$  of size  $n$  and we construct a full binary tree of height  $n$ , as follows. The root of the tree is encoded by a visible unary relation  $N_0$ , which is initialized with a single value  $y_0$ . To encode the nodes of the tree at the lower levels, we use a series of hidden unary relations  $N_1, \dots, N_n$ . Similarly, we encode the edges at each level of the binary tree using a series of hidden binary relations  $E_{1,\text{left}}, E_{1,\text{right}}, \dots, E_{n,\text{left}}, E_{n,\text{right}}$ . The corresponding constraints are easily stated:

$$\begin{array}{ll}
 N_0(y) \rightarrow \exists y' E_{1,\text{left}}(y, y') & E_{1,\text{left}}(y, y') \rightarrow N_1(y') \\
 N_0(y) \rightarrow \exists y' E_{1,\text{right}}(y, y') & E_{1,\text{right}}(y, y') \rightarrow N_1(y') \\
 \vdots & \vdots \\
 N_{n-1}(y) \rightarrow \exists y' E_{n,\text{left}}(y, y') & E_{n,\text{left}}(y, y') \rightarrow N_n(y') \\
 N_{n-1}(y) \rightarrow \exists y' E_{n,\text{right}}(y, y') & E_{n,\text{right}}(y, y') \rightarrow N_n(y') .
 \end{array}$$

Note that, in a universal instance that satisfies the above constraints (e.g., an instance obtained from the chase procedure), every cell of the tape can be identified with a unique element of the relation  $N_n$ . The paths in the full binary tree are naturally ordered lexicographically, and so are the cells. Later, we need to access this ordering and, in particular, we need to write a UCQ that checks whether two cells are adjacent according to the ordering. For this, we need to make the encoding a bit redundant. We first introduce two visible unary relations,  $D_{\text{left}}$  and  $D_{\text{right}}$ , that are instantiated, respectively, with the singleton  $\{\text{left}\}$  and the singleton  $\{\text{right}\}$ . The values  $\text{left}$  and  $\text{right}$  that appear in these

sets are used to encode whether a certain node of the binary tree is a left or a right successor of its parent. First, this information is collected in a series of hidden binary relations  $D_{1,\text{left}}, D_{1,\text{right}}, \dots, D_{n,\text{left}}, D_{n,\text{right}}$  (two relations for each level of the tree, except for the top level). Then, every pair of relations  $D_{i,\text{left}}$  and  $D_{i,\text{right}}$  is unioned into a new relation  $D_i$ , which is also hidden. The objective is to have a collection of facts of the form  $D_i(y, d)$ , where  $y$  is a node,  $i$  is its level, and  $d$  is either **left** or **right** depending on whether  $y$  is a left or right successor of its parent. It is easy to see that this objective is achieved when chasing the following constraints:

$$\begin{array}{ll}
E_{1,\text{left}}(y) \rightarrow \exists d D_{1,\text{left}}(y, d) & E_{1,\text{right}}(y) \rightarrow \exists d D_{1,\text{right}}(y, d) \\
\vdots & \vdots \\
E_{n,\text{left}}(y) \rightarrow \exists d D_{n,\text{left}}(y, d) & E_{n,\text{right}}(y) \rightarrow \exists d D_{n,\text{right}}(y, d) \\
D_{i,\text{left}}(y, d) \rightarrow D_{\text{left}}(d) & D_{i,\text{right}}(y, d) \rightarrow D_{\text{right}}(d) \\
D_{i,\text{left}}(y, d) \rightarrow D_i(y, d) & D_{i,\text{right}}(y, d) \rightarrow D_i(y, d) .
\end{array}$$

To explain how we can take advantage of the above redundant encoding, we give beforehand the formula that checks whether two cells  $y$  and  $y'$  are adjacent. The idea is to find the first level  $0 \leq i < n$  in the tree where the access paths to the leaves  $y$  and  $y'$  branches off; after this level, the access path for  $y$  must continue following the direction **right**, and the access path for  $y'$  must continue following the direction **left**. The formula is the disjunction over all  $i = 0, \dots, n-1$  of the following CQs:

$$\begin{aligned}
Q_{\text{adj},i}(y_n, y'_n) = & \exists y_0 \dots y_{n-1} y'_0 \dots y'_{n-1} d_1 \dots d_i d d' \\
& N_0(y_0) \wedge \bigwedge_{0 < j \leq i} (E_j(y_{j-1}, y_j) \wedge D_j(y_j, d_j)) \wedge \\
& N_0(y'_0) \wedge \bigwedge_{0 < j \leq i} (E_j(y'_{j-1}, y'_j) \wedge D_j(y'_j, d_j)) \wedge \\
& D_{\text{right}}(d) \wedge \bigwedge_{i < j \leq n} (E_j(y_{j-1}, y_j) \wedge D_j(y_j, d)) \wedge \\
& D_{\text{left}}(d') \wedge \bigwedge_{i < j \leq n} (E_j(y'_{j-1}, y'_j) \wedge D_j(y'_j, d')) .
\end{aligned}$$

It is not difficult to see that the formula correctly defines those pairs of cells that are adjacent in the tape, under the usual assumption that the instance is generated by chasing the constraints, namely, that the instance is universal.

Now that we constructed a tape of exponential length and we know how to check adjacency of its cells, we proceed as in the proof of Theorem 7.

We begin by encoding a single configuration of  $M$ . Intuitively, this is done by creating a copy of the full binary tree and expanding it with the identifier of the configuration and the content of the cells (i.e. the cell values). For this, we introduce a series of hidden binary relations  $N_0^C, \dots, N_n^C$ , a series of hidden ternary relations  $E_{1,\text{left}}^C, E_{1,\text{right}}^C, \dots, E_{n,\text{left}}^C, E_{n,\text{right}}^C$ , and an additional hidden ternary relation  $C$ . The content for the relations  $N_0^C, \dots, N_n^C$

and  $E_{1,\text{left}}^C, E_{1,\text{right}}^C, \dots, E_{n,\text{left}}^C, E_{n,\text{right}}^C$  will be obtained by copying the content of  $N_0, \dots, N_n, E_{1,\text{left}}, E_{1,\text{right}}, \dots, E_{n,\text{left}}, E_{n,\text{right}}$  and by annotating it with the identifier  $x$  of the configuration. Formally, this is done through the follows constraints:

$$\begin{array}{llll}
N_0^C(x, y) & \rightarrow & N_0(y) & \\
N_1^C(x, y) & \rightarrow & N_1(y) & E_1^C(x, y, y') \rightarrow E_1(y, y') \\
& \vdots & & \vdots \\
N_n^C(x, y) & \rightarrow & N_n(y) & E_n^C(x, y, y') \rightarrow E_n(y, y') \\
\\ 
N_0^C(y) & \rightarrow & \exists y' E_{1,\text{left}}^C(y, y') & E_{1,\text{left}}^C(y, y') \rightarrow N_1(y') \\
N_0^C(y) & \rightarrow & \exists y' E_{1,\text{right}}^C(y, y') & E_{1,\text{right}}^C(y, y') \rightarrow N_1(y') \\
& \vdots & & \vdots \\
N_{n-1}^C(y) & \rightarrow & \exists y' E_{n,\text{left}}^C(y, y') & E_{n,\text{left}}^C(y, y') \rightarrow N_n^C(y') \\
N_{n-1}^C(y) & \rightarrow & \exists y' E_{n,\text{right}}^C(y, y') & E_{n,\text{right}}^C(y, y') \rightarrow N_n^C(y') .
\end{array}$$

The ternary relation  $C$  is used instead to represent the values of the tape cells. Intuitively, a fact of the form  $C(x, y, z)$  indicates that, in the configuration identified by  $x$ , the cell  $y$  has value  $z$ . As usual (cf. proof of Theorem 7), we define cell values as elements from a visible unary relation  $V = (\Sigma \times \Sigma \times Q \times \Sigma) \uplus (\Sigma \times Q \times \Sigma) \uplus (\Sigma \times \Sigma \times Q) \uplus \Sigma$ , where  $\Sigma$  is the alphabet of the Turing machine and  $Q$  is the set of its control states. We recall that if a cell has value  $(a, b, q, c)$ , this means that its content is  $b$ , the control state of  $M$  is  $q$ , the head is on this cell, and the neighbouring cells have labels  $a$  and  $c$ . Analogous semantics are given for the values of the form  $(a, q, b)$  (resp.,  $(b, c, q)$ ), which must be associated with cells that are immediately to the right (resp., to the left) of the head of the Turing machine. We enforce the following constraints over  $N_n^C$ ,  $C$ , and  $V$ :

$$N_n^C(x, y) \rightarrow \exists z C(x, y, z) \quad C(x, y, z) \rightarrow V(z) .$$

We now turn to the encoding of the computation tree of  $M$ . This is almost the same as in the proof of Theorem 7. We introduce a visible unary relation  $C_0$ , which contains the identifier of the initial configuration, and three hidden binary relations  $S^\exists$ ,  $S_1^\forall$ , and  $S_2^\forall$ . A fact of the form  $S^\exists(x, x')$  (resp.,  $S_1^\forall(x, x_1)$ ,  $S_1^\forall(x, x_1)$ ) represents a transition from an existential (resp., universal) configuration  $x$  to a universal (resp., existential) configuration  $x'$  (resp.,  $x_1, x_2$ ). We then enforce the following constraints:

$$\begin{array}{llll}
C_0(x) & \rightarrow & \exists x' S^\exists(x, x') & S^\exists(x, x') \rightarrow \exists x_1 S_1^\forall(x', x_1) \\
S^\exists(x, x') & \rightarrow & \exists y' N_0^C(x, y) & S^\exists(x, x') \rightarrow \exists x_2 S_2^\forall(x', x_2) \\
S_1^\forall(x, x_1) & \rightarrow & \exists y' N_0^C(x, y) & S_1^\forall(x, x_1) \rightarrow \exists x' S^\exists(x_1, x') \\
S_2^\forall(x, x_2) & \rightarrow & \exists y' N_0^C(x, y) & S_2^\forall(x, x_2) \rightarrow \exists x' S^\exists(x_2, x') .
\end{array}$$

Now, we turn to describing how to detect badly formed encodings of the computation tree of  $M$ . We introduce the visible relations  $\text{Err}_{\perp,0}$ ,  $\text{Err}_{\perp}$ ,  $\text{Err}_C$ ,

$\text{Err}_{S^\exists}$ ,  $\text{Err}_{S_1^\forall}$ , and  $\text{Err}_{S_2^\forall}$ , whose instances are defined exactly as in the proof of Theorem 7.

- Recall that the relation  $\text{Err}_{\perp,0}$  contains all pairs in  $V \times V$  but  $(z_0, z_1)$ , where  $z_0$  is the value  $(\perp, \perp, q_0)$  of the first cell of the initial configuration and  $z_1 = (\perp, \perp, q_0, \perp)$  is the value of the second cell of the initial configuration. We can detect whether the first two cells of the initial configuration are badly formed using the CQ

$$\begin{aligned} Q_{\perp,0} = & \exists x y_0 \dots y_n y'_n d d' z z' \\ & C_0(x) \wedge D_{\text{left}}(d) \wedge D_{\text{right}}(d') \wedge N_0^C(x, y_0) \wedge \\ & \bigwedge_{0 < i \leq n} (E_i^C(y_{i-1}, y_i) \wedge D_i(y_i, d)) \wedge (E_i^C(y_{n-1}, y'_n) \wedge D_n(y_{n-1}, d')) \wedge \\ & C(x, y_n, z) \wedge C(x, y'_n, z') \wedge \text{Err}_{\perp,0}(z, z') . \end{aligned}$$

Similarly, the relation  $\text{Err}_1$  contains those pairs of values that cannot occur in two adjacent cells of the initial configuration. We can detect whether two adjacent cells of the initial configuration contain wrong values by a disjunction of CQs that are built up using the previous formulas  $Q_{\text{adj},i}$ , for  $i = 0, \dots, n-1$ . The resulting UCQ is

$$\begin{aligned} Q_{\perp} = & \bigvee_{0 \leq i < n} \exists x y y' z z' C_0(x) \wedge N_n^C(x, y) \wedge N_n^C(x, y') \wedge Q_{\text{adj},i}(y, y') \wedge \\ & C(x, y, z) \wedge C(x, y', z') \wedge \text{Err}_{\perp}(z, z') . \end{aligned}$$

- The relation  $\text{Err}_C$  contains those pairs of cell values  $(z, z')$  that cannot be adjacent, in any configuration of  $M$ . A violation of the adjacency constraint can be detected by the following UCQ:

$$\begin{aligned} Q_C = & \bigvee_{0 \leq i < n} \exists x y y' z z' N_n^C(x, y) \wedge N_n^C(x, y') \wedge Q_{\text{adj},i}(y, y') \wedge \\ & C(x, y, z) \wedge C(x, y', z') \wedge \text{Err}_C(z, z') . \end{aligned}$$

- For the violations that involve values associated with the same position of the tape in two subsequent configurations, we use a disjunction of the following CQs:

$$\begin{aligned} Q_{S^\exists} &= \exists x x' y z z' S^\exists(x, x') \wedge C(x, y, z) \wedge C(x', y, z') \wedge \text{Err}_{S^\exists}(z, z') \\ Q_{S_1^\forall} &= \exists x x_1 y z z' S_1^\forall(x, x_1) \wedge C(x, y, z) \wedge C(x_1, y, z') \wedge \text{Err}_{S_1^\forall}(z, z') \\ Q_{S_2^\forall} &= \exists x x_2 y z z' S_2^\forall(x, x_2) \wedge C(x, y, z) \wedge C(x_2, y, z') \wedge \text{Err}_{S_2^\forall}(z, z') . \end{aligned}$$

As usual, we can further check whether the Turing machine  $M$  reaches the rejecting state  $q_{\text{rej}}$  along some path of its computation tree. This is done by the CQ

$$Q_{\text{rej}} = \exists x y z C(x, y, z) \wedge V_{\text{rej}}(z) .$$

where  $V_{\text{rej}}$  is an additional visible relation that contains all cell values of the form  $(a, b, q_{\text{rej}}, c)$ , for some  $a, b, c \in \Sigma$ .

Let  $Q$  be the disjunction of all the previous CQs and let  $\mathcal{V}$  be the instance that captures the intended semantics of the visible relations  $N_0$ ,  $D_{\text{left}}$ ,  $D_{\text{right}}$ ,  $V$ ,  $\text{Err}_{\perp,0}$ ,  $\text{Err}_{\perp}$ ,  $\text{Err}_C$ ,  $\text{Err}_{S^{\exists}}$ ,  $\text{Err}_{S_1^{\forall}}$ , and  $\text{Err}_{S_2^{\forall}}$ . To conclude, we argue along the same lines of the proof of Theorem 7 that  $M$  has a successful computation tree iff  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{false}$ .  $\square$

### 3.2 Existence problems

We now turn to the schema-level problem  $\exists\text{PQI}$ . Let  $\mathcal{V}_{\{a\}}$  be a fixed instance for the visible part of a schema  $\mathbf{S}$  whose domain contains the single value  $a$  and whose visible relations are singleton relations of the form  $\{(a, \dots, a)\}$ . For certain constraint languages, we will show that, whenever  $\exists\text{PQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$ , then the witnessing instance can be taken to be  $\mathcal{V}_{\{a\}}$ . This can be viewed as an extension of the “critical instance” method which has been applied previously to chase termination problems: Proposition 3.7 of Marnette and Geerts [MG10] states a related result for disjunctive TGDs in isolation; Gogacz and Marcinowski [GM14] call such an instance a “well of positivity”. The following shows that the technique applies to TGDs and EGDs without constants.

**Theorem 10.** *For every Boolean UCQ  $Q$  and every set  $\mathcal{C}$  of TGDs and EGDs without constants,  $\exists\text{PQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  iff  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}}) = \text{true}$ .*

We prove the above theorem first for constraints consisting only of TGDs without constants; then we will show how to generalize the proof in the additional presence of EGDs without constants. First of all, recall that, by introducing additional invisible relations, we can assume, without loss of generality, that all TGDs have exactly one atom in the right-hand side.

Next, we introduce a variant of the chase procedure that returns a collection of instances (not necessarily finite). As for the classical chase, the procedure receives as input a relational schema  $\mathbf{S}$ , some constraints  $\mathcal{C}$ , and an initial instance  $\mathcal{F}_0$  for the schema  $\mathbf{S}$ , which does not need to satisfy the constraints in  $\mathcal{C}$ . The procedure chases the constraints starting from the instance  $\mathcal{F}_0$ , guaranteeing at the same time that the visible relations of the constructed instances agree with  $\mathcal{F}_0$ . This variant of the chase will be used to prove Theorem 10, as well as other results related to the  $\exists\text{NQI}$  problem.

Formally, the procedure builds a *chase tree* of instances, starting with the singleton tree consisting of the input  $\mathbf{S}$ -instance  $\mathcal{F}_0$  and extending the tree by repeatedly applying the following steps. It chooses an instance  $K$  at some leaf of the current tree, a dependency  $R_1(\bar{x}_1) \wedge \dots \wedge R_m(\bar{x}_m) \rightarrow \exists \bar{y} S(\bar{z})$ , where  $\bar{z}$  is a sequence of (possibly repeated) variables from  $\bar{x}_1, \dots, \bar{x}_m, \bar{y}$ , and a homomorphism  $f$  that maps  $R_1(\bar{x}_1), \dots, R_m(\bar{x}_m)$  to some facts in  $K$ . Then, the procedure constructs a new instance from  $K$  by adding the fact  $S(f'(\bar{z}))$ , where  $f'$  is an extension of  $f$  that maps, in an injective way, the existentially quantified

variables in  $\bar{y}$  to some fresh null values (this can be seen as a classical chase step). Immediately after, and only when the relation  $S$  is visible, the procedure replaces the instance  $K' = K \cup \{S(f'(\bar{z}))\}$  with copies of it of the form  $g(K')$  such that  $\text{Visible}(g(K')) = \text{Visible}(\mathcal{F}_0)$ , for all possible homomorphisms  $g$  that map the variables  $\bar{z}$  to some values in the active domain  $\{a_1, \dots, a_n\}$  of the visible instance  $\text{Visible}(\mathcal{F}_0)$  (this can be seen as a chase step for disjunctive EGDs of the form  $S(\bar{z}) \rightarrow \bar{z}(i) = a_1 \vee \dots \vee \bar{z}(i) = a_n$ ). The resulting instances  $g(K')$  are then appended as new children of  $K$  in the tree-shaped collection. In the special case where there are no homomorphisms  $g$  such that  $\text{Visible}(g(K')) = \text{Visible}(\mathcal{F}_0)$ , we append a “dummy instance”  $\perp$  as a child of  $K$ : this is used to represent the fact that the chase step from  $K$  led to an inconsistency (the dummy node will never be extended during the subsequent chase steps). If  $S$  is not visible, then the instance  $K'$  is simply appended as a new child of  $K$ .

This process continues iteratively using a strategy that is “fair”, namely, that guarantees that whenever a dependency is applicable in a node on a maximal path of the chase tree, then it will be fired at some node (possibly later) on that same maximal path (unless the path ends with  $\perp$ ). In the limit, the process generates a possibly infinite tree-shaped collection of instances. It remains to complete the collection with “limits” in order to guarantee that the constraints are satisfied. Consider any infinite path  $K_0, K_1, \dots$  in the tree (if there are any). It follows from the construction of the chase tree that the instances on the path form a chain of homomorphic embeddings  $K_0 \xrightarrow{h_0} K_1 \xrightarrow{h_1} \dots$ . Such chains of homomorphic embeddings admit a natural notion of limit, which we denote by  $\lim_{n \in \mathbb{N}} K_n$ . We omit the details of this construction here, which can be found, for instance, in [CK90]. The limit instance  $\lim_{n \in \mathbb{N}} K_n$  satisfies the constraints  $\mathcal{C}$ . We denote by  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{F}_0)$  the collection of all non-dummy instances that occur at the leaves of the chase tree, plus all limit instances of the form  $\lim_{n \in \mathbb{N}} K_n$ , where  $K_0, K_1, \dots$  is an infinite path in the chase tree. This is well-defined only once the ordering of steps is chosen, but for the results below, which order is chosen will not matter, so we abuse notation by referring to  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{F}_0)$  as a single object.

It is clear that every instance in  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{F}_0)$  satisfies the constraints in  $\mathcal{C}$  and, in addition, agrees with  $\mathcal{F}_0$  on the visible part of the schema. Below, we prove that  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{F}_0)$  satisfies the following universal property:

**Lemma 11.** *Let  $\mathcal{F}_0$  be an instance of a schema  $\mathbf{S}$  and let  $\mathcal{F}$  be another instance over the same schema that contains  $\mathcal{F}_0$ , agrees with  $\mathcal{F}_0$  on the visible part (i.e.  $\text{Visible}(\mathcal{F}) = \text{Visible}(\mathcal{F}_0)$ ), and satisfies a set  $\mathcal{C}$  of TGDs without constants. Then, there exist an instance  $K \in \text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{F}_0)$  and a homomorphism from  $K$  to  $\mathcal{F}$ .*

*Proof.* We consider the chase tree for  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{F}_0)$  and, based on the full instance  $\mathcal{F}$ , we identify inside this chase tree a suitable path  $K_0, K_1, \dots$  and a corresponding sequence of homomorphisms  $h_0, h_1, \dots$  such that, for all  $n \in \mathbb{N}$ ,  $h_n$  maps  $K_n$  to  $\mathcal{F}$ . Once these sequences are defined, the lemma will follow easily by letting  $K = \lim_{n \in \mathbb{N}} K_n$  and  $h = \lim_{n \in \mathbb{N}} h_n$ , that is,  $h(\bar{a}) = \bar{b}$  if  $h_n(\bar{a}) = \bar{b}$  for all but finitely many  $n \in \mathbb{N}$ .

The base step is easy, as we simply let  $K_0$  be the initial instance  $\mathcal{F}_0$ , which appears at the root of the chase tree, and let  $h_0$  be the identity. As for the inductive step, suppose that  $K_n$  and  $h_n$  are defined for some step  $n$ , and suppose that  $R_1(\bar{x}_1) \wedge \dots \wedge R_m(\bar{x}_m) \rightarrow \exists \bar{y} S(\bar{z})$  is the dependency that is applied at node  $K_n$ , where  $\bar{z}$  is a sequence of variables from  $\bar{x}_1, \dots, \bar{x}_m, \bar{y}$ . Let  $R_1(f(\bar{x}_1)), \dots, R_m(f(\bar{x}_m))$  be the facts in the instance  $K_n$  that have triggered the chase step, where  $f$  is an homomorphism from the variables in  $\bar{x}_1, \dots, \bar{x}_m$  to the domain of  $K_n$ . Since  $\mathcal{F}$  satisfies the same dependency and contains the facts  $R_1(h_n(f(\bar{x}_1))), \dots, R_m(h_n(f(\bar{x}_m)))$ , it must also contain a fact of the form  $S(h'(f'(\bar{z})))$ , where  $f'$  is the extension of  $f$  that is the identity on the existentially quantified variables  $\bar{y}$  and  $h'$  is some extension of  $h_n$  that maps the variables  $\bar{y}$  to some values in the domain of  $\mathcal{F}$ .

Now, to choose the next instance  $K_{n+1}$ , we distinguish two cases, depending on whether  $S$  is visible or not. If  $S$  is not visible, then we know that the chase step appends a single instance  $K' = K_n \cup \{S(h'(\bar{z}))\}$  as a child of  $K_n$ ; accordingly, we let  $K_{n+1} = K'$  and  $h_{n+1} = h' \circ f'$ . Otherwise, if  $S$  is visible, then we observe that  $h'$  is a homomorphism from  $K' = K_n \cup \{S(h'(\bar{z}))\}$  to  $\mathcal{F}$ . In particular,  $h'$  maps the variables  $\bar{z}$  to some values in the active domain of the visible part  $\text{Visible}(\mathcal{F}_0)$  and hence  $h'(K')$  agrees with  $\mathcal{F}_0$  on the visible part of the schema. This implies that the chase step adds at least the instance  $h'(K')$  as a child of  $K_n$ . Accordingly, we can define  $K_{n+1} = h'(K')$  and  $h_{n+1} = f'$ . Given the above constructions, it is easy to see that the homomorphism  $h_{n+1}$  maps  $K_{n+1}$  to  $\mathcal{F}$ .

Proceeding in this way, we either arrive at a leaf, in which case we are done, or we obtain an infinite path of the chase tree  $K_0 \xrightarrow{h_0} K_1 \xrightarrow{h_1} \dots$ , with homomorphisms  $h'_i : K_i \rightarrow \mathcal{F}$ , such that  $h_i \cdot h'_{i+1}$  extends  $h'_i$ , for all  $i \in \mathbb{N}$ . It can be shown that, then, the limit  $\lim_{n \in \mathbb{N}} K_n$  also homomorphically maps to  $\mathcal{F}$ .  $\square$

**Proposition 12.** *If  $Q$  is a Boolean UCQ,  $\mathcal{C}$  is a set of TGDs without constants over a schema  $\mathbf{S}$ , and  $\mathcal{V}$  is a visible instance, then  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$  iff every instance  $K$  in  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V})$  satisfies  $Q$ .*

*Proof.* Suppose that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$  and recall that every instance in  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V})$  satisfies the constraints in  $\mathcal{C}$  and agrees with  $\mathcal{V}$  on the visible part. In particular, this means that every instance in  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V})$  satisfies the query  $Q$ .

Conversely, suppose that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{false}$ . This means that there is an  $\mathbf{S}$ -instance  $\mathcal{F}$  that has  $\mathcal{V}$  as visible part, satisfies the constraints in  $\mathcal{C}$ , but not the query  $Q$ . By Lemma 11, letting  $\mathcal{F}_0 = \mathcal{V}$ , we get an instance  $K \in \text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V})$  and a homomorphism from  $K$  to  $\mathcal{F}$ . Since  $Q$  is preserved under homomorphisms,  $K$  does not satisfy  $Q$ .  $\square$

Next, we recall that the visible instance  $\mathcal{V}_{\{a\}}$  is constructed over a singleton active domain and the constraints  $\mathcal{C}$  have no constants. This implies that there are no disjunctive choices to perform while chasing the constraints starting from  $\mathcal{V}_{\{a\}}$ . Moreover, it is easy to see that this chase always succeeds. That is, it



returns a collection  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$  with exactly one instance – in particular,  $\mathcal{V}_{\{a\}}$  is a realizable instance. By a slight abuse of notation, we denote by  $\text{chase}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$  the unique instance in the collection  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$ .

**Lemma 13.** *If  $\mathcal{C}$  is a set of TGDs without constants over a schema  $\mathbf{S}$  and  $\mathcal{V}$  is an instance of the visible part of  $\mathbf{S}$ , then every instance  $K \in \text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V})$  maps homomorphically to  $\text{chase}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V}_{\{a\}})$ , that is,  $h(K) \subseteq \text{chase}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V}_{\{a\}})$  for some homomorphism  $h$ .*

*Proof.* Recall that the instances in  $\text{Chases}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V})$  are either leaves or limits of infinite paths of the chase tree. Below, we prove that every instance  $K$  in the chase tree for  $\text{Chases}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V})$  maps to  $\text{chase}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V}_{\{a\}})$  via some homomorphism  $h$ . In addition, we ensure that, if  $K'$  is a descendant of  $K$  in the same chase tree, then the corresponding homomorphism  $h'$  is obtained by composing some homomorphism with an extension of  $h$ . This way of constructing homomorphisms is compatible with limits in the following sense: if  $h_0, h_1, \dots$  are homomorphisms mapping instances  $K_0, K_1, \dots$  along an infinite path of the chase tree, then there is a homomorphism  $\lim_{n \in \mathbb{N}} h_n$  that maps the limit instance  $\lim_{n \in \mathbb{N}} K_n$  to  $\mathcal{F}$ .

For the base case of the induction, we consider the initial instance  $\mathcal{V}$  at the root of the chase tree, which clearly maps homomorphically to  $\mathcal{V}_{\{a\}}$ . For the inductive case, we consider an instance  $K$  in the chase tree and suppose that it maps to  $\text{chase}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V}_{\{a\}})$  via a homomorphism  $h$ . We also consider an instance  $K'$  that is a child of  $K$  and is obtained by chasing some dependency  $R_1(\bar{x}_1) \wedge \dots \wedge R_m(\bar{x}_m) \rightarrow \exists \bar{y} S(\bar{z})$ , where  $\bar{z}$  is a sequence of variables from  $\bar{x}_1, \dots, \bar{x}_m, \bar{y}$ . This means that there exist two homomorphisms  $f$  and  $g$  such that

1.  $f$  maps the variables  $\bar{x}_1, \dots, \bar{x}_m$  to some values in  $K$  and maps injectively the variables  $\bar{y}$  to fresh values;
2.  $g$  either maps  $f(\bar{z})$  to values in the active domain of  $\mathcal{V}$  or is the identity on  $f(\bar{z})$ , depending on whether  $S$  is visible or not;
3.  $R_j(f(\bar{x}_j)) \in K$  for all  $1 \leq j \leq m$ ;
4.  $K' = g(K \cup \{S(f(\bar{z}))\})$ .

Note that  $h$  maps each fact  $R_j(f(\bar{x}_j))$  in  $K$  to  $R_j(h(f(\bar{x}_j)))$  in  $\text{chase}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V}_{\{a\}})$ . Since  $\text{chase}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V}_{\{a\}})$  satisfies the chased dependency, it must also contain a fact of the form  $S(h'(f(\bar{z})))$ , where  $h'$  is a homomorphism that extends  $h$  on the fresh values  $f(\bar{y})$ . Moreover, if  $S$  is visible, then  $h'$  maps all values  $f(\bar{z})$  to the same value  $a$ , which is the only element of the active domain of  $\mathcal{V}_{\{a\}}$ .

We can now define a homomorphism that maps the instance  $K' = g(K \cup \{S(f(\bar{z}))\})$  to  $\text{chase}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V}_{\{a\}})$ . If  $S$  is not visible, then we recall that  $g$  is the identity on  $f(\bar{z})$ , and hence  $h'$  already maps  $K' = g(K \cup \{S(f(\bar{z}))\}) = K \cup \{S(f(\bar{z}))\}$  to  $\text{chase}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V}_{\{a\}})$ . Otherwise, if  $S$  is visible, then we recall that  $g$  maps  $f(\bar{z})$  to values in the active domain of  $\mathcal{V}$ , we let  $g'$  be the function that maps all values of the active domain of  $\mathcal{V}$  to  $a$ , and finally we define  $h'' = h' \circ g'$ . In this way  $h''$  maps  $K' = g(K \cup \{S(f(\bar{z}))\})$  to  $\text{chase}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V}_{\{a\}})$ .  $\square$

Now that we established the key lemmas, we can easily reduce the existence problem to an instance-based problem (recall that for the moment we assume that the constraints consist only of TGDs):

*Proof of Theorem 10.* Clearly,  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}}) = \text{true}$  implies  $\exists \text{PQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$ . For the converse direction, suppose that  $\exists \text{PQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$ . This implies the existence of a realizable instance  $\mathcal{V}$  such that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ . By Proposition 12, every instance in  $\text{Chases}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V})$  satisfies the query  $Q$ . Moreover, by Lemma 13, every instance in  $\text{Chases}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V})$  maps homomorphically to  $\text{chase}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V}_{\{a\}})$ . Hence  $\text{chase}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V}_{\{a\}})$  also satisfies  $Q$ . By applying Proposition 12 again, we conclude that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}}) = \text{true}$ .

Finally, the second statement of the theorem follows from the fact that the previous proofs are independent of the assumption that relational instances are finite.  $\square$

Now, we explain how to generalize the proof of Theorem 10 to constraints consisting of both TGDs and EGDs (still without constants). This can be done by modifying the chase procedure for  $\text{Chases}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V})$  so as to take into account also the EGDs in  $\mathcal{C}$  that can be triggered on the instances that emerge in the chase tree. Formally, chasing an EGD of the form  $R_1(\bar{x}_1) \wedge \dots \wedge R_m(\bar{x}_m) \rightarrow x = x'$ , where  $x, x'$  are two variables from  $\bar{x}_1, \dots, \bar{x}_m$ , amounts at applying a suitable homomorphism that identifies the two values  $h(x)$  and  $h(x')$  whenever the facts  $R_1(h(\bar{x}_1)), \dots, R_m(h(\bar{x}_m))$  belong to the instance under consideration. Note that this operation leads to a failure (i.e. a dummy instance) when  $h(x)$  and  $h(x')$  are distinct values from the active domain of the visible part  $\mathcal{V}$ .

With this new definition of  $\text{Chases}_{\text{vis}}(\mathbf{S}, \mathcal{C}, \mathcal{V})$  at hand, the proofs of Lemma 11 and Lemma 13 do not pose particular problems, as one just needs to handle the standard case of an EGD dependency. Finally, the proof of Theorem 10 directly uses Proposition 12 and Lemma 13 as black boxes, and so carries over without any modification.

It is worth remarking that, by pairing Theorem 10 with the upper bound and the finite controllability for instance-level problems (Theorem 4), one immediately obtains the following:

**Corollary 14.**  *$\exists \text{PQI}(Q, \mathcal{C}, \mathbf{S})$  with  $Q$  ranging over Boolean UCQs and  $\mathcal{C}$  over sets of frontier-guarded TGDs without constants, is decidable in 2EXPTIME, and is finitely controllable.*

In contrast, we show that adding disjunctions or constants to the constraints leads to undecidability. We first prove this in the case where the constraints have disjunctions. This shows that the interaction of disjunctive linear TGDs and linear EGDs (implicit in the visibility assumption) causes the “critical instance” reduction to fail.

**Theorem 15.** *The problem  $\exists \text{PQI}(Q, \mathcal{C}, \mathbf{S})$  is undecidable as  $Q$  ranges over Boolean UCQs and  $\mathcal{C}$  over sets of disjunctive linear TGDs.*

The proof uses a technique that will be exploited for many of our schema-level undecidability arguments. We will reduce the existence of a tiling to the  $\exists\text{PQI}$  problem. The tiling itself will correspond to the visible instance that has a  $\text{PQI}$ . The invisible relations will store “challenges” to the correctness of the tiling. The UCQ  $Q$  will have disjuncts that return `true` exactly when the challenge to correctness is passed. There will be challenges to the labelling of adjacent cells, challenges to the correctness of the initial tile, and challenges to the correct shape of the adjacency relationship – that is, challenges that the tiling is really grid-like. A correct tiling corresponds to every challenge being passed, and thus corresponds to a visible instance where every extension satisfies  $Q$ . The undecidability argument also applies to the “unrestricted version” of  $\exists\text{PQI}$ , in which both quantifications over instances consider arbitrary instances. This will also be true for all other undecidability results in this work, which always concern the schema-level problems.

*Proof.* For simplicity, we deal with the “unrestricted variant” of the problem, which asks if there is an arbitrary instance of the visible schema such that every superinstance satisfying the constraints also satisfies  $Q$ . We comment below on how to modify for finite instances.

We reduce the problem of tiling the infinite grid, which is known to be undecidable, to the problem  $\exists\text{PQI}$ . Recall that an instance of the tiling problem consists of a finite set  $T$  of available tiles, a set of horizontal and vertical constraints, given by relations  $H, V \subseteq T \times T$ , and an initial tile  $t_\perp \in T$  for the lower-left corner. The problem consists of deciding whether there is a tiling function  $f : \mathbb{N} \times \mathbb{N} \rightarrow T$  such that

1.  $f(0, 0) = t_\perp$ ,
2.  $(f(i, j), f(i + 1, j)) \in H$  for all  $i, j \in \mathbb{N}$ ,
3.  $(f(i, j), f(i, j + 1)) \in V$  for all  $i, j \in \mathbb{N}$ .

Given an instance  $(T, H, V, t_0)$  of the tiling problem, we show how to construct a schema  $\mathbf{S}$ , a query  $Q$ , and a set of disjunctive IDs over  $\mathbf{S}$  such that  $\exists\text{PQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  if and only if there is a tiling function for  $(T, H, V, t_0)$ .

The idea is to enforce suitable constraints and query in such a way that the visible instance that witnesses  $\exists\text{PQI}$  represents a candidate tiling, and invisible instances represent challenges to the correctness of the tiling. We use attributes to denote cells of the grid and we use two visible binary relations  $E_H$  and  $E_V$  to represent the horizontal and vertical edges of the grid. We also introduce a unary visible relation  $U_t$  for each tile  $t \in T$  in order to represent a candidate tiling function on the grid.

We begin by enforcing the existence of an initial node with the associated tile  $t_\perp$ . For this, we introduce another visible relation `linit`, of arity 0, and linear TGD

$$\text{linit} \rightarrow \exists x U_{t_\perp}(x) .$$

It is also easy to require that every node is connected to at least another node in the relation  $E_H$  (resp.,  $E_V$ ), and that the latter node has an associated tile that satisfies the horizontal constraints  $H$  (resp., the vertical constraints  $V$ ). To do so we use the disjunctive linear TGDs

$$U_t(x) \rightarrow \exists y E_H(x, y) \wedge \bigvee_{(t, u) \in H} U_u(y) \quad (\text{for all tiles } t \in T)$$

$$U_t(x) \rightarrow \exists z E_V(x, z) \wedge \bigvee_{(t, v) \in V} U_v(z) \quad (\text{for all tiles } t \in T)$$

We now turn to explain how to enforce a grid structure on the relations  $E_H$  and  $E_V$ , and to guarantee that each node has exactly one tile associated with it. Of course, we cannot directly use disjunctive TGDs in order to guarantee that  $E_H$  and  $E_V$  correctly represent the horizontal and vertical edges of the grid. However, we can introduce additional hidden relations that make it possible to mark certain nodes so as to expose the possible violations. We first show how expose violations to the fact that the horizontal edge relation is a function. The idea is to select nodes in  $E_H$  in order to challenge functionality. Formally, the horizontal challenge is captured by a hidden ternary relation **HChallenge**, by the linear TGDs

$$\begin{aligned} \text{Init} &\rightarrow \exists x y y' \text{HChallenge}(x, y, y') \\ \text{HChallenge}(x, y, y') &\rightarrow E_H(x, y) \wedge E_H(x, y') \end{aligned}$$

and by the CQ

$$Q_H = \exists x y \text{HChallenge}(x, y, y) .$$

Note that if the visible fact **Init** is present and the relation  $E_H$  correctly describes the horizontal edges of the grid, then the above query  $Q_H$  is necessarily satisfied by any instance of **HChallenge** that satisfies the above constraints: the only way to give a non-empty instance for **HChallenge**<sub>funct</sub> is to use triples of the form  $(x, y, y)$ . Conversely, if the relation  $E_H$  is not a function, namely, if there exist nodes  $x, y, y'$  such that  $(x, y), (x, y') \in E_H$  and  $y \neq y'$ , then the singleton instance  $\{(x, y, y')\}$  for the hidden relation **HChallenge**<sub>funct</sub> will satisfy the associated constraint and violate the query  $Q_H$ . Note that we do not require that the relation  $E_H$  is injective (this could be still done, but is not necessary for the reduction). Similarly, we can use a hidden relation **VChallenge** and analogous constraints and query  $Q_V$  in order to challenge the functionality of  $E_V$ .

In the same way, we can challenge the confluence of the relations  $E_H$  and  $E_V$ . For this, we introduce a hidden relation **CChallenge** of arity 5, which is associated with the constraints

$$\begin{aligned} \text{Init} &\rightarrow \exists x y z w w' \text{CChallenge}(x, y, z, w, w') \\ \text{CChallenge}(x, y, z, w, w') &\rightarrow E_H(x, y) \wedge E_V(x, z) \wedge E_V(y, w) \wedge E_H(z, w') \end{aligned}$$

and the CQ

$$Q_C = \exists x y z w \text{CChallenge}(x, y, z, w, w) .$$

As before, we can argue that there is a positive query implication for  $Q_C$  iff the horizontal and vertical edge relations are confluent, that is,  $(x, w) \in E_H \circ E_V$  and  $(x, w') \in E_V \circ E_H$  imply  $w = w'$ .

We need to ensure that there does not exist a node  $n$  labeled with two tiles, which means that there does not exist two relations  $U_t$  and  $U_{t'}$  such that  $n$  is not in  $U_t(\mathcal{V})$  and  $U_{t'}(\mathcal{V})$ .

For that we add the two following constraints, where  $A$  and  $B$  are hidden relations

$$\text{Init} \rightarrow \exists x A(x) \vee B(x)$$

$$B(x) \rightarrow \bigvee_{t \neq t'} (U_t(x) \wedge U_{t'}(x))$$

Finally, there exists a CQ

$$Q_A = \exists x A(x)$$

Now that we described all the visible and hidden relations of the schema  $\mathbf{S}$ , and the associated constraints  $\mathcal{C}$ , we define the query for the  $\exists\text{PQI}$  problem as the conjunction of the atom  $\text{Init}$  and all previous UCQs (for this we distribute the disjunctions and existential quantifications over the conjunctions):

$$Q = \text{Init} \wedge Q_A \wedge Q_H \wedge Q_V \wedge Q_C \wedge Q_T.$$

It remains to show that  $\exists\text{PQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  iff there is a correct tiling of the infinite grid, namely, a function  $f : \mathbb{N} \times \mathbb{N} \rightarrow T$  that satisfies the conditions 1), 2), and 3) above.

Suppose there exists a correct tiling  $f : \mathbb{N} \times \mathbb{N} \rightarrow T$ . We construct the visible instance  $\mathcal{V}$  that contains the fact  $\text{Init}$  and the relations  $E_H$ ,  $E_V$ , and  $U_t$  with the intended semantics:  $E_H = \{((i, j), (i + 1, j)) \mid i, j \in \mathbb{N}\}$ ,  $E_V = \{((i, j), (i, j + 1)) \mid i, j \in \mathbb{N}\}$ , and  $U_t = \{(i, j) \mid f(i, j) = t\}$  for all  $t \in T$ . Since no error can be exposed on the relations  $E_H$ ,  $E_V$ , and  $U_t$ , no matter how we construct a full instance  $\mathcal{F}$  that agrees with  $\mathcal{V}$  on the visible part and satisfies the constraints in  $\mathcal{C}$ , we will have that  $\mathcal{F}$  satisfies all the components of the query  $Q$ , other than  $Q_A$ . In addition, in any such  $\mathcal{F}$ ,  $B$  must be empty, since otherwise tiling predicates for distinct tiles would overlap, which is not the case. Since  $\text{Init}$  holds, we can conclude via the first constraint above that  $Q_A$  must hold.

Conversely, suppose that  $\exists\text{PQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  and let  $\mathcal{V}$  be the witnessing visible instance. Clearly,  $\mathcal{V}$  contains the fact  $\text{Init}$  (otherwise, the query would be immediately violated) and it does not contain  $W$ , otherwise the query  $Q_A$  would be violated. We can use the content of  $\mathcal{V}$  and the knowledge that  $\exists\text{PQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  to inductively construct a correct tiling of the infinite grid. More precisely, by the first constraint in  $\mathcal{C}$ , we know that  $\mathcal{V}$  contains the fact  $U_{t_1}(x)$ , for some node  $x$ . Accordingly, we define  $i_x = 0$ ,  $j_x = 0$ , and  $f(i_x, j_x) = t_1$ . For the induction step, suppose that  $f(i_x, j_x)$  is defined for a node  $x$  with the associated coordinates  $i_x$  and  $j_x$ . The constraints in  $\mathcal{C}$  enforce the existence of two cells  $y$  and  $z$  and two tiles  $t$  and  $t'$  for which the following facts are in the visible instance:  $E_H(x, y)$ ,  $E_V(x, z)$ ,  $U_t(y)$ , and  $U_{t'}(z)$ . Accordingly, we let  $i_y = i_x + 1$ ,

$j_y = j_x$ ,  $i_z = i_x$ ,  $j_z = j_y + 1$ ,  $f(i_y, j_y) = t$ , and  $f(i_z, j_z) = t'$ . By the initial constraints, we know that the tiles associated with the new cells  $(i_y, j_y)$  and  $(i_z, j_z)$  are consistent with the tile in  $(i_x, j_x)$  and with the horizontal and vertical constraints  $H$  and  $V$ . We now argue that there is a unique choice for the nodes  $y$  and  $z$ . Indeed, suppose this is not the case; for instance, suppose that there exist two distinct nodes  $y, y'$  that are connected to  $x$  via  $E_H$ . Then, we could construct a full instance in which the relation **HChallenge** contains the single triple  $(x, y, y')$ . This will immediately violate the CQ  $Q_H$ , and hence  $Q$ . Similar arguments apply to the vertical successor  $z$ .

We now argue that there are unique choices for the tile  $t$  associated with a node  $y$ . Suppose not. Then we can set  $A$  to empty,  $B$  to all nodes having multiple tiles. The constraints are satisfied, while  $Q_A$  fails, hence we have violated the assumption that we have a PQL.

Finally, we can argue along the same lines that, during the next steps of the induction, the  $E_V$ -successor of  $y$  and the  $E_H$ -successor of  $z$  coincide. The above properties are sufficient to conclude that the constructed function  $f$  is a correct tiling of the infinite grid.

The variant for finite instances is done by observing that the same reduction produces a periodic grid, which can be represented as a finite instance.  $\square$

Perhaps even more surprisingly, we show that *disjunction can be simulated using constants (under UNA)*. The proof, works by applying the technique of “coding Boolean operations and truth values in the schema” which has been used to eliminate the need for disjunction in hardness proofs in several past works (e.g. [GP03]).

**Proposition 16.** *There is a polynomial time reduction from  $\exists\text{PQL}(Q, \mathcal{C}, \mathbf{S})$ , where  $Q$  ranges over Boolean UCQs and  $\mathcal{C}$  over sets of disjunctive linear TGDs, to  $\exists\text{PQL}(Q', \mathcal{C}', \mathbf{S}')$ , where  $Q'$  ranges over Boolean UCQs and  $\mathcal{C}'$  over sets of linear TGDs (with constants).*

*Proof.* We transform the schema  $\mathbf{S}$  to a new schema  $\mathbf{S}'$  as follows. For every visible (resp., hidden) relation  $R$  of  $\mathbf{S}$  of arity  $k$ , we add to  $\mathbf{S}'$  a corresponding visible (resp., hidden) relation  $R'$  of arity  $k + 1$ . The idea is that the additional attribute of  $R'$  represents a truth value, i.e. either the constant 0 or the constant 1, which indicates the presence of a tuple in the original relation  $R$ . For example, the fact  $R'(\bar{a}, 1)$  indicates the presence of the tuple  $\bar{a}$  in the relation  $R$ . We can then simulate the disjunctions in the constraints of  $\mathcal{C}$  by using conjunctions and an appropriate look-up table, which we denote by **Or**. Formally, we introduce three additional relations **Or**, **Check**, and **Init**, of arities 2, 1, and 0, respectively, and we let **Or** and **Init** be visible and **Check** be hidden in  $\mathbf{S}'$ . Then, for every disjunctive linear TGD in  $\mathcal{C}$  of the form

$$R(\bar{x}) \rightarrow \exists \bar{y} S(\bar{z}) \vee T(\bar{z}')$$

we add to  $\mathcal{C}'$  the linear TGD with constants

$$R'(\bar{x}, 1) \rightarrow \exists \bar{y} b_1 b_2 S'(\bar{z}, b_1) \wedge T(\bar{z}', b_2) \wedge \text{Or}(b_1, b_2) .$$

We further add to  $\mathcal{C}'$  the following constraints:

$$\begin{aligned}\text{Init} &\rightarrow \text{Or}(0, 1) \wedge \text{Or}(1, 0) \wedge \text{Or}(1, 1) \\ \text{Init} &\rightarrow \exists b_1 b_2 \text{Or}(b_1, b_2) \wedge \text{Check}(b_1) \wedge \text{Check}(b_2) .\end{aligned}$$

Finally, we transform every CQ of  $Q$  of the form  $\exists \bar{y} S(\bar{y})$  to a corresponding CQ of  $Q'$  of the form

$$\exists \bar{y} S'(\bar{y}, 1) \wedge \text{Check}(1) \wedge \text{Init}$$

Note that if needed, we can even rewrite the CQ above so as to avoid constants: we introduce another hidden unary relation **One** and the constraint  $\text{Init} \rightarrow \text{One}(1)$ , and we replace the conjunct  $\text{Check}(1)$  with  $\exists b \text{Check}(b) \wedge \text{One}(b)$ . Below, we prove that  $\exists \text{PQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  iff  $\exists \text{PQI}(Q', \mathcal{C}', \mathbf{S}') = \text{true}$ .

For the easier direction, we consider a realizable  $\mathbf{S}_v$ -instance  $\mathcal{V}$  such that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ . We can easily transform  $\mathcal{V}$  into a realizable  $\mathbf{S}'_v$ -instance  $\mathcal{V}'$  that satisfies  $\text{PQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}') = \text{true}$ . For this it suffices to copy the content of the visible relations of  $\mathcal{V}$  into  $\mathcal{V}'$ , by properly expanding the tuples with the constant 1, and then adding the facts  $\text{Init}$ ,  $\text{Or}(0, 1)$ ,  $\text{Or}(1, 0)$ , and  $\text{Or}(1, 1)$ .

As for the converse direction, we consider a realizable  $\mathbf{S}'_v$ -instance  $\mathcal{V}'$  such that  $\text{PQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}') = \text{true}$ . By the definition of  $Q'$  it is clear that  $\mathcal{V}'$  contains the fact  $\text{Init}$ , and hence also the facts  $\text{Or}(0, 1)$ ,  $\text{Or}(1, 0)$ , and  $\text{Or}(1, 1)$ . Now, if we knew that the relation  $\text{Or}$  contains no other tuples besides  $(0, 1)$ ,  $(1, 0)$ , and  $(1, 1)$  – that is, for every fact  $\text{Or}(b_1, b_2)$  in  $\mathcal{V}'$ , we have  $b_1 = 1$  or  $b_2 = 1$  – then we could easily transform  $\mathcal{V}'$  into a realizable  $\mathbf{S}_v$ -instance  $\mathcal{V}$  that satisfies  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ . For this we simply select the facts  $R'(\bar{a}, 1)$  in  $\mathcal{V}'$ , where  $R$  is a visible relation of  $\mathbf{S}$ , and project away the constant 1.

It remains to show that the relation  $\text{Or}$  of  $\mathcal{V}'$  indeed contains no other tuples besides  $(0, 1)$ ,  $(1, 0)$ , and  $(1, 1)$ . For the sake of contradiction, suppose that  $\mathcal{V}'$  contains a fact of the form  $\text{Or}(b_1, b_2)$ , with  $b_1 \neq 1$  and  $b_2 \neq 1$ . Since  $\mathcal{V}'$  is realizable, there is a full  $\mathbf{S}'$ -instance  $\mathcal{F}'$  such that  $\mathcal{F}' \models \mathcal{C}'$  and  $\text{Visible}(\mathcal{F}') = \mathcal{V}'$ . Note that  $\mathcal{F}'$  may satisfy  $Q'$  and, in particular, the conjunct  $\text{Check}(1)$ . However, removing the single fact  $\text{Check}(1)$  from  $\mathcal{F}'$  gives a new instance  $\mathcal{F}''$  that still satisfies the constraints in  $\mathcal{C}'$ , agrees with  $\mathcal{F}'$  on the visible part, and violates the query  $Q'$ . This contradicts the fact that  $\text{PQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}') = \text{true}$ .  $\square$

From the previous two results we immediately see that the addition of (distinct) constants leads to undecidability:

**Corollary 17.** *The problem  $\exists \text{PQI}(Q, \mathcal{C}, \mathbf{S})$  is undecidable as  $Q$  ranges over Boolean CQs and  $\mathcal{C}$  over sets of linear TGDs (with constants).*

We now turn to analysing how the complexity scales with less powerful constraints, e.g. linear TGDs without constants. As before, we reduce  $\exists \text{PQI}(Q, \mathcal{C}, \mathbf{S})$  to  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$ . We can then reuse some ideas from [JK84] to solve the latter problem in polynomial space:

**Theorem 18.** *The problem  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$  as  $Q$  ranges over Boolean UCQs and  $\mathcal{C}$  over sets of linear TGDs without constants, is in PSPACE, and the same is true for  $\exists\text{PQI}(Q, \mathcal{C}, \mathbf{S})$ .*

*Proof.* By Proposition 12,  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}}) = \text{true}$  is equivalent to checking that there is a homomorphism  $h$  from  $\text{CanonDB}(Q_i)$  of some CQ  $Q_i$  of  $Q$  to the instance  $\text{chase}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$ . We can easily guess in NP a CQ  $Q_i$  of  $Q$ , some homomorphism  $h$  from  $\text{CanonDB}(Q_i)$ , and the corresponding image  $I$  of  $\text{CanonDB}(Q_i)$  under  $h$ . Then, it remains to decide whether  $I$  is contained in  $\text{chase}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$ . Below, we explain how to decide this in polynomial space.

Recall that the instance  $\text{chase}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$  is obtained as the limit of a series of operations that consist of alternatively adding new facts according to the TGDs in  $\mathcal{C}$  and identifying the values that appear in some visible relation with the constant  $a$ . Note that the second type of operation may also affect tuples that belong to hidden relations (this happens when the values are shared with facts in the visible instance). Also note that the affected tuples could have been inferred during previous steps of the chase. Nonetheless, at the exact moment when a new fact  $R(b_1, \dots, b_k)$  is inferred by chasing a linear TGD, we can detect whether a certain value  $b_i$  needs to be eventually identified with the constant  $a$ , and in this case we can safely replace the fact  $R(b_1, \dots, b_k)$  with  $R(b_1, \dots, b_{i-1}, a, b_{i+1}, \dots, b_k)$ . More precisely, to decide whether the  $i$ -th attribute of  $R(\bar{b})$  needs to be instantiated with the constant  $a$ , we test whether  $\mathcal{C}$  entails a dependency of the form  $R(\bar{x}) \rightarrow \exists \bar{y} S(\bar{z})$ , where  $\bar{x}$  is a sequence of (possibly repeated) variables that has the same equality type as  $\bar{b}$  (i.e.  $\bar{x}(j) = \bar{x}(j')$  iff  $\bar{b}(j) = \bar{b}(j')$ ),  $S$  is a visible relation,  $\bar{z}$  is a sequence of variables among  $\bar{x}, \bar{y}$ , and  $\bar{x}(i) = \bar{z}(j)$  for some  $1 \leq j \leq |\bar{z}|$ . Note that the above entailment can be rephrased as a containment problem between two CQs – i.e.  $R(\bar{x})$  and  $\exists \bar{y} S(\bar{z})$  – under a given set of linear TGDs  $\mathcal{C}$ , and we know from [JK84] that the latter problem is in PSPACE. We also observe that, in order to discover all the values in  $R(\bar{b})$  that need to be identified with the constant  $a$ , it is not sufficient to execute the above analysis only once on each position  $1 \leq i \leq \text{ar}(R)$ , as identifying some values with the constant  $a$  may change the equality type of the fact and thus trigger new dependencies from  $\mathcal{C}$  (notably, this may happen when the linear TGDs are not IDs). We thus repeat the above analysis on all positions of  $R$  and until the corresponding equality type stabilizes – this can be still be done in polynomial space. After this, we add the resulting fact to the chase.

What we have just described is an alternative construction of  $\text{chase}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$  in which every chase step can be done using a PSPACE sub-procedure. We omit the tedious details showing that this alternative construction gives the same result, in the limit, as the version of the chase that we introduced at the beginning of Section 3.2 (the arguments are similar to the proof of Lemma 11).

Below, we explain how to adapt the techniques from [JK84] to this alternative variant of the chase, in order to decide whether the homomorphic image  $I$  of some CQ of  $Q$  is contained in  $\text{chase}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$ . For this, it is convenient



to think of  $\text{chase}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$  as a directed graph, where the nodes represent the facts in  $\text{chase}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$  and the edges describe the inference steps that derive new facts from existing ones and constraints in  $\mathcal{C}$  – note that, because the constraints are linear TGDs, each inference step depends on at most one fact. In particular, the nodes of this graph that have no incoming edge (we call them *roots*) are precisely the facts from the instance  $\mathcal{V}_{\{a\}}$ , and all the other nodes are reachable from some root. Moreover, by the previous arguments, one can check in polynomial space whether an edge exists between two given nodes.

Now, we focus on the minimal set of edges that connect all the facts of  $I$  to some roots in the graph. The graph restricted to this set of edges is a forest, namely, every node in it has at most one incoming edge. Moreover, the height of this forest is at most exponential in  $|I|$ , and each level in it contains at most  $|I|$  nodes. Thus, the restricted graph can be explored by a non-deterministic polynomial-space algorithm that guesses the nodes at a level on the basis of the nodes at the previous level and the linear TGDs in  $\mathcal{C}$ . The algorithm terminates successfully once it has visited all the facts in  $I$ , witnessing that  $I$  is contained in  $\text{chase}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \mathcal{V}_{\{a\}})$ . Otherwise, the computation is rejected after seeing exponentially many levels.  $\square$

We can derive matching lower bounds by reducing Open-World Query Answering to  $\exists\text{PQI}$ :

**Proposition 19.** *For any class of constraints containing linear TGDs, OWQ reduces to  $\exists\text{PQI}$ .*

*Proof.* Let  $Q$  be a query,  $\mathcal{C}$  a set of constraints over a schema  $\mathbf{S}$ , and  $\mathcal{F}$  an instance of the schema  $\mathbf{S}$ . We show how to reduce the Open-World Query Answering problem for  $Q$ ,  $\mathcal{C}$ ,  $\mathbf{S}$ , and  $\mathcal{F}$  to a problem  $\exists\text{PQI}(Q', \mathcal{C}', \mathbf{S}')$ . The idea is to create a copy of the instance  $\mathcal{F}$  in the hidden part of the schema, which can be then extended arbitrarily.

Formally, we let the transformed schema  $\mathbf{S}'$  consist of all the relations in  $\mathbf{S}$ , which are assumed to be hidden, plus an additional visible relation **Good** of arity 0. We then introduce a variable  $y_b$  for each value in the active domain of  $\mathcal{F}$ , and we let  $\mathcal{C}'$  contain all the constraints from  $\mathcal{C}$ , plus a constraint of the form  $\text{Good} \rightarrow \exists \bar{y} Q_{\mathcal{F}}$ , where  $\bar{y}$  contains one variable  $y_b$  for each value  $b$  in the active domain of  $\mathcal{F}$  and  $Q_{\mathcal{F}}$  is the conjunction of the atoms of the form  $A(y_{b_1}, \dots, y_{b_k})$ , for all facts  $A(b_1, \dots, b_k)$  in  $\mathcal{F}$ . Note that the visible instance  $\mathcal{V}_{\text{Good}}$  that contains the atom **Good** is realizable, since it can be completed (using the chase) to an  $\mathbf{S}'$ -instance  $\mathcal{F}'$  that satisfies the constraints  $\mathcal{C}'$ . Let  $Q' = Q \wedge \text{Good}$ . We claim that  $\exists\text{PQI}(Q', \mathcal{C}', \mathbf{S}') = \text{true}$  if and only if  $Q$  is certain with respect to  $\mathcal{C}$  on  $\mathcal{F}$ . In one direction, suppose  $\exists\text{PQI}(Q', \mathcal{C}', \mathbf{S}') = \text{true}$  holds. The witness visible instance having **PQI** can only be the instance  $\mathcal{V}_{\text{Good}}$ . Consider an instance  $\mathcal{F}'$  containing all facts of  $\mathcal{F}$  and satisfying the original constraints. By setting **Good** to true in  $\mathcal{F}'$ , we have an instance satisfying  $\mathcal{C}'$ , and since  $\mathcal{V}_{\text{Good}}$  has a **PQI** then we know that this instance must satisfy  $Q'$  and hence  $Q$ . Thus  $Q$  is certain with respect to  $\mathcal{C}$  on  $\mathcal{F}$  as required. Conversely, suppose  $Q$  is certain with respect to  $\mathcal{C}$  on  $\mathcal{F}$ . Letting  $\mathcal{C}_{\mathcal{F}}$  be the chase of  $\mathcal{F}$  with respect to  $\mathcal{C}$ , we see that  $\mathcal{C}_{\mathcal{F}}$  satisfies  $Q$ . We

will show there is a PQI for  $Q', \mathcal{C}', \mathbf{S}$  on  $\mathcal{V}_{\text{Good}}$ . Thus fix an instance  $\mathcal{F}'$  where **Good** and  $\mathcal{C}'$  holds. The additional constraint implies that  $\mathcal{F}'$  contains the image of  $\mathcal{F}$  under some homomorphism  $h$ . But  $h$  extends to a homomorphism of  $C_{\mathcal{F}}$  into  $\mathcal{F}'$ . Thus  $\mathcal{F}'$  satisfies  $Q$ , and therefore satisfies  $Q'$ . Thus there is a PQI on  $\mathcal{V}_{\text{Good}}$  as required.

Thus we have reduced the Open-World Query Answering problem for  $Q, \mathcal{C}$ , and  $\mathbf{S}$  to the problem  $\exists\text{PQI}(Q', \mathcal{C}', \mathbf{S}')$ .  $\square$

From this and existing lower bounds on the Open-World Query Answering ([CFP84] coupled with a reduction from implication to OWQ for linear TGDs, [CGK13] for FGTGDs), we see that the prior upper bounds from Theorem 18 and Corollary 14 are tight:

**Corollary 20.** *The problem  $\exists\text{PQI}(Q, \mathcal{C}, \mathbf{S})$ , where  $Q$  ranges over CQs and  $\mathcal{C}$  over sets of linear TGDs, is PSPACE-hard.*

**Corollary 21.** *The problem  $\exists\text{PQI}(Q, \mathcal{C}, \mathbf{S})$ , where  $Q$  ranges over CQs and  $\mathcal{C}$  over sets of FGTGDs without constants, is 2EXPTIME-hard.*

### 3.3 Summary for Positive Query Implication

The main results on positive query implication are highlighted in the table below.

	PQI Data	PQI Combined	$\exists\text{PQI}$
NoConst	EXPTIME-cmp	2EXPTIME-cmp	PSPACE-cmp
Linear TGD	Thm 5/Thm 7	Thm 4/Thm 9	Thm 18/Cor 20
NoConst	EXPTIME-cmp	2EXPTIME-cmp	2EXPTIME
FGTGD	Thm 5/Thm 7	Thm 4/Thm 9	Cor 14/Cor 21
NoConst Disj.	EXPTIME-cmp	2EXPTIME-cmp	undecidable
Linear TGD	Thm 5/Thm 7	Thm 4/Thm 9	Thm 15
Linear TGD & FGTGD & GNFO	EXPTIME-cmp Thm 5/Thm 7	2EXPTIME-cmp Thm 4/Thm 9	undecidable Cor 17

## 4 Negative Query Implication

### 4.1 Instance-level problems

Here we analyze the complexity of the problem  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$ . As in the positive case, we begin with an upper bound that holds for a very rich class of constraints, which go far beyond referential constraints (and FGTGDs).

**Theorem 22.** *The problem  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$ , as  $Q$  ranges over Boolean UCQ and  $\mathcal{C}$  over sets of GNFO constraints, has 2EXPTIME combined complexity, EXPTIME data complexity, and it is finitely controllable.*

*Proof.* As in the positive case, we reduce to unsatisfiability of a GNFO formula. We use a variation of the same formula:

$$\phi_{Q,\mathcal{C},\mathbf{S},\mathcal{V}}^{\text{NQItoGNF}} = Q \wedge \mathcal{C} \wedge \bigwedge_{R \in \mathbf{S}_v} \left( \bigwedge_{R(\bar{a}) \in \mathcal{V}} R(\bar{a}) \wedge \forall \bar{x} \left( R(\bar{x}) \rightarrow \bigvee_{R(\bar{a}) \in \mathcal{V}} \bar{x} = \bar{a} \right) \right)$$

The data complexity analysis is as in Theorem 5, since the formulas agree on the part that varies with the instance.  $\square$

We can show that this bound is tight if the class of constraints is rich enough. This follows from our lower bound for positive query implication problems, since we can show that NQI is at least as difficult as PQI for powerful constraints.

**Theorem 23.** *For any class of constraints that include connected FGTGDs,  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  reduces in polynomial time to  $\text{NQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}')$ . When  $Q, \mathcal{C}, \mathbf{S}$  are fixed in the input to this reduction, then  $Q', \mathcal{C}', \mathbf{S}'$  are fixed in the output. Thus, for these classes of constraints, the lower bounds for combined and data complexity given in Theorems 7 and 9 apply to negative query implications as well.*

*Proof.* We first provide a reduction that works with any class of constraints allowing arbitrary conjunctions in the left-hand sides (e.g. frontier-guarded TGDs). Subsequently, we show how to modify the constructions in order to preserve connectedness.

The schema  $\mathbf{S}'$  is obtained by copying both the visible and the hidden relations from  $\mathbf{S}$  and by adding the following relations: a visible relation **Error** of arity 0 and a hidden relation **Good** of arity 0. The constraints  $\mathcal{C}'$  will contain the same constraints from  $\mathcal{C}$ , plus one frontier-guarded TGD of the form

$$Q_i(\bar{y}) \wedge \text{Good} \rightarrow \text{Error}$$

for each CQ of  $Q$  of the form  $\exists \bar{y} Q_i(\bar{y})$ . Finally, the query and the visible instance for NSB are defined as follows:  $Q' = \text{Good}$  and  $\mathcal{V}' = \mathcal{V}$ .

We now verify that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{false}$  iff  $\text{NQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}') = \text{false}$ . Suppose that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{false}$ , namely, that there is an  $\mathbf{S}$ -instance  $\mathcal{F}$  such that  $\mathcal{F} \not\models Q$ ,  $\mathcal{F} \models \mathcal{C}$ , and  $\text{Visible}(\mathcal{F}) = \mathcal{V}$ . Let  $\mathcal{F}'$  be the  $\mathbf{S}'$ -instance obtained from  $\mathcal{F}$  by adding the single hidden fact **Good**. Clearly,  $\mathcal{F}'$  satisfies the query  $Q'$  and also the constraints in  $\mathcal{C}'$ ; in particular, it satisfies every constraint  $Q_i(\bar{y}) \wedge \text{Good} \rightarrow \text{Error}$  because  $\mathcal{F}$  violates every disjunct  $\exists \bar{y} Q_i$  of  $Q$ . Hence, we have  $\text{NQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}') = \text{false}$ . Conversely, suppose that  $\text{NQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}') = \text{false}$ , namely, that there is an  $\mathbf{S}'$ -instance  $\mathcal{F}'$  such that  $\mathcal{F}' \models Q'$ ,  $\mathcal{F}' \models \mathcal{C}'$ , and  $\text{Visible}(\mathcal{F}') = \mathcal{V}'$ . By copying the content of  $\mathcal{F}'$  for those relations belong to the schema  $\mathbf{S}$ , we obtain an  $\mathbf{S}$ -instance  $\mathcal{F}$  that satisfies the constraints  $\mathcal{C}$ . Moreover, because  $\mathcal{F}'$  contains the fact **Good** but not the fact **Error**,  $\mathcal{F}'$  violates every conjunct  $\exists \bar{y} Q_i(\bar{y})$  of  $Q$ , and so  $\mathcal{F}$  does. This shows that  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{false}$ .

We observe that the constraints in the above reduction use left-hand sides that are not connected. In order to preserve connectedness, it is sufficient to modify the above constructions by adding a dummy variable that is shared among all atoms. More precisely, we expand the relations of the schema  $\mathbf{S}$  and the relation **Good** with a new attribute, and we introduce a new visible relation **Check** of arity 1. The dummy variable will be used to enforce connectedness in the left-hand sides, and the relation **Check** will gather all the values associated with the dummy attribute. Using the visible instance, we can also check that the relation **Check** contains exactly one value. The constraints are thus modified as follows. Every constraint  $R_1(\bar{x}_1) \wedge \dots \wedge R_m(\bar{x}_m) \rightarrow \exists \bar{y} S(\bar{z})$  in  $\mathcal{C}'$  is transformed into  $R_1(\bar{x}_1, w) \wedge \dots \wedge R_m(\bar{x}_m, w) \rightarrow \exists \bar{y} S(\bar{z}, w)$ . In particular, note that the constraint  $Q_i(\bar{y}) \wedge \text{Good} \rightarrow \text{Error}$  becomes  $Q_i(\bar{y}, w) \wedge \text{Good}(w) \rightarrow \text{Error}(w)$ , which is now a connected frontier-guarded TGD. Furthermore, for every relation  $R(\bar{x})$  in  $\mathbf{S}$ , we add the constraint

$$R(\bar{x}, w) \rightarrow \text{Check}(w)$$

and we do the same for the relation **Good**:

$$\text{Good}(w) \rightarrow \text{Check}(w) .$$

Finally, the query is transformed into  $Q' = \exists w \text{Good}(w)$  and the visible instance  $\mathcal{V}'$  is expanded with a fresh dummy value  $a$  on the additional attribute and with the visible fact **Check**( $a$ ).  $\square$

As mentioned in the body, from the above reduction and from Theorems 7 and 9, we get the following hardness results for instance-based NSB.

**Corollary 24.** *There are a Boolean UCQ  $Q$  and a set  $\mathcal{C}$  of IDs over a schema  $\mathbf{S}$  for which the problem  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  is EXPTIME-hard in data complexity (that is, as  $\mathcal{V}$  varies over instances).*

**Corollary 25.** *The problem  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$ , as  $\mathcal{C}$  ranges over sets of connected frontier-guarded TGDs,  $\mathbf{S}$  over schemas,  $Q$  over conjunctive queries and  $\mathcal{V}$  over instances, is 2EXPTIME-hard.*

Thus far, the negative query implication results have been similar to the positive ones. We will now show a strong contrast in the case of IDs and linear TGDs. Recall that the PQI problems were highly intractable even with for fixed schema, query, and constraints. We begin by showing that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  can be solved easily by looking only at full instances that agree with  $\mathcal{V}$  on the visible part and whose active domains are the same as that of  $\mathcal{V}$ :

**Definition 26.** *The problem  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  is said to be active domain controllable if it is equivalent to asking that for every instance  $\mathcal{F}$  over the active domain of  $\mathcal{V}$ , if  $\mathcal{F}$  satisfies  $\mathcal{C}$  and  $\mathcal{V} = \text{Visible}(\mathcal{F})$ , then  $Q(\mathcal{F}) = \text{false}$ .*

It is clear that the the problem  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  is simpler when it is active domain controllable, as in this case we could guess a full instance  $\mathcal{F}$  over the active domain of  $\mathcal{V}$  and then reduce the problem to checking whether  $Q$  holds on  $\mathcal{F}$ .

We give a simple argument that  $\text{NQL}$  under IDs is active domain controllable. Let  $\mathcal{C}$  be a set of IDs over a schema  $\mathbf{S}$ ,  $Q$  be a UCQ, and  $\mathcal{V}$  be a visible instance such that  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{false}$ . This means that there is a full instance  $\mathcal{F}$  such that  $\mathcal{F} \models \mathcal{C}$ ,  $\text{Visible}(\mathcal{F}) = \mathcal{V}$ , and  $\mathcal{F} \models Q$ . Now take any value  $a \in \text{adom}(\mathcal{V})$  (i.e. in the active domain of  $\mathcal{V}$ ) and let  $h$  be the homomorphism that is the identity over  $\text{adom}(\mathcal{V})$  and maps any other value from  $\text{adom}(\mathcal{F}) \setminus \text{adom}(\mathcal{V})$  to  $a$ . Since, the constraints  $\mathcal{C}$  are IDs (in particular, since the left-hand side atoms do not have repeated occurrences of the same variable), we know that  $h(J) \models \mathcal{C}$ . Similarly, we have  $h(J) \models Q$ . Hence,  $h(J)$  is an instance over the active domain of  $\mathcal{V}$  that equally witnesses  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{false}$ .

The following example shows that linear TGDs are not always active domain controllable.

**Example 3.** Let  $\mathbf{S}$  be the schema with a hidden relation  $R$  of arity 2, with two visible relations  $S, T$  of arities 1, 0, respectively, and with the constraints:

$$R(x, y) \rightarrow S(x) \qquad R(x, x) \rightarrow T.$$

Note that the constraints are linear TGDs and they are even full – no existential quantifiers on the right. The conjunctive query is  $Q = \exists x y R(x, y)$ . Further let the visible instance  $\mathcal{V}$  consists of the single fact  $S(a)$ . Clearly, every full instance  $\mathcal{F}$  over the active domain  $\{a\}$  that satisfies both  $\mathcal{C}$  and  $Q$  must also contain the facts  $R(a, a)$  and  $T$ , and so such an instance cannot agree with  $\mathcal{V}$  in the visible part. On the other hand, the instance that contains the facts  $S(a)$  and  $R(a, b)$ , for a fresh value  $b$ , satisfies both  $\mathcal{C}$  and  $Q$  and moreover agrees with  $\mathcal{V}$ . This shows that  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  is not active domain controllable.

Despite the above example, we show that we can still transform any schema with linear TGDs (which may include constants) and any query for an  $\text{NQL}$  problem so as to enforce active domain controllability. Furthermore, we can do so while preserving the visible instance of the problem:

**Theorem 27.** Given a schema  $\mathbf{S}$ , a set  $\mathcal{C}$  of linear TGDs (possibly including constants), and a UCQ  $Q$ , one can construct in exponential time a new schema  $\mathbf{S}'$ , a set  $\mathcal{C}'$  of linear TGDs (with constants), and a UCQ  $Q'$  such that  $\mathbf{S}_v = \mathbf{S}'_v$  and, for all instances  $\mathcal{V}$  over  $\mathbf{S}_v$ :

1.  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{NQL}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V})$ ,
2.  $\text{NQL}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V})$  is active domain controllable,
3. the number of constraints of  $\mathcal{C}'$  is exponential in that of  $\mathcal{C}$ , but each constraint of  $\mathcal{C}'$  is polynomial in the maximum size of the constraints of  $\mathcal{C}$ , and similarly for the number and size of CQs in  $Q$  and  $Q'$ .

*Proof.* The main idea of the transformation is to select those attributes in a relation of  $\mathbf{S}$  that are going to be instantiated with values from the active domain  $\text{adom}(\mathcal{V})$  of  $\mathcal{V}$ . Since we do not know in advance which attributes need to

be selected (this depends in particular on the instance  $\mathcal{V}$ ), we introduce several copies of the hidden relations of  $\mathbf{S}$ , one for each possible choice of set of selected attributes. Moreover, in order to be able to correctly reconstruct some witnessing instances, we need to recall the equality relationships enforced on those attributes that are *not* selected. This is necessary since there might exist linear TGDs with repeated variables in the left-hand side that are activated only when the corresponding attributes carry the same value. For similar reasons (e.g., presence of constants in the left-hand side of a TGD), we must also recall the equalities enforced between unselected attributes and constants outside  $\text{adom}(\mathcal{V})$ .

Formally, we define an *equality pattern* over a set  $I$  of attributes as a sequence of (possibly repeated) variables and constants indexed by  $I$ . Like the atoms of a TGD, equality patterns can be used to recall equality relationships between attributes and constants, but not inequalities. We generically denote equality patterns by  $\bar{u}, \bar{v}$ , etc. Moreover, we compare equality patterns up to variable renaming, that is, we write  $\bar{u} \approx \bar{v}$  whenever  $\bar{u} = h(\bar{v})$  holds for some injective function  $h$  from variables to variables. Similarly, we write  $\bar{u} \leq \bar{v}$  whenever  $\bar{u} = h(\bar{v})$  for some (possibly non-injective) function  $h$  from variables to variables and constants. For example,  $\bar{u} \leq \bar{v}$  for  $\bar{u} = xyctxy$  and  $\bar{v} = xyzwy$ . For each  $\approx$ -equivalence class, we fix, once and for all, an equality pattern that acts as a representative of the class.

The transformed schema  $\mathbf{S}'$  contains all the visible relations of the original schema  $\mathbf{S}$ , plus one relation  $R_{I,\bar{u}}$  of arity  $|I|$  for each hidden relation  $R$  in  $\mathbf{S}$ , each set  $I \subseteq \{1, \dots, \text{ar}(R)\}$ , and each representative  $\bar{u}$  of an  $\approx$ -equivalence class over  $\tilde{I} = \{1, \dots, \text{ar}(R)\} \setminus I$ . Even if we do not enforce it explicitly, the attributes that are selected in a copy  $R_{I,\bar{u}}$  of  $R$  are meant to contain only values ranging over the active domain of  $\mathcal{V}$ . For convenience, we also denote the visible relations in the transformed schema  $\mathbf{S}'$  by  $R_{I,\bar{u}}$ , where  $I$  is assumed to be the full set of attributes  $\{1, \dots, \text{ar}(R)\}$  and  $\bar{u}$  is the trivial equality pattern over the empty set.

Accordingly, every constraint in  $\mathcal{C}$  involving some relations  $R$  and  $S$  is translated to several analogous constraints in  $\mathcal{C}'$  involving relations of the form  $R_{I,\bar{u}}$  and  $S_{J,\bar{v}}$ . More precisely, we consider every linear TGD in  $\mathcal{C}$  of the form

$$R(\bar{x}) \rightarrow \exists \bar{y} S(\bar{z})$$

where  $\bar{x}$  is a sequence of variables and constants,  $\bar{y}$  is a sequence of variables, and  $\bar{z}$  is a sequence of variables and constants from  $\bar{x}$  and  $\bar{y}$ . We also consider every possible pair of relations  $R_{I,\bar{u}}$  and  $S_{J,\bar{v}}$  in  $\mathbf{S}'$ . Then, for each of the above choices, we add to  $\mathcal{C}'$  the linear TGD

$$R_{I,\bar{u}}(\bar{x}|I) \rightarrow \exists \bar{y} S_{J,\bar{v}}(\bar{z}|J)$$

provided that the following conditions hold:

1.  $I$  contains at least the positions of the left-hand side  $R(\bar{x})$  that are associated with constants from  $\text{adom}(\mathcal{V})$ ,

2.  $J$  contains at least the positions of the left-hand side  $S(\bar{z})$  that are associated with constants from  $\text{adom}(\mathcal{V})$ ,
  3. if  $\bar{x}(i) = \bar{z}(j)$ , then  $i \in I$  iff  $j \in J$ ,
  4.  $\bar{x}|\bar{I} \leq \bar{u}$ , where  $\bar{I} = \{1, \dots, \text{ar}(R)\} \setminus I$ ,
  5.  $\bar{z}|\bar{J} \leq \bar{v}$ , where  $\bar{J} = \{1, \dots, \text{ar}(S)\} \setminus J$
- (of course, if  $R$  is visible then there is a single choice for  $I$  and  $\bar{u}$ , and similarly for  $J$  and  $\bar{v}$  when  $S$  is visible).

The transformation of the query  $Q$  is similar: for all CQs

$$\exists \bar{y} S_1(\bar{z}_1) \wedge \dots \wedge S_n(\bar{z}_n)$$

in  $Q$  and for all sequences of relations  $S_{1,J_1,\bar{v}_1}, \dots, S_{n,J_n,\bar{v}_n}$  in  $\mathbf{S}'$  such that  $\bar{z}_1|\bar{J}_1 \leq \bar{v}_1, \dots, \bar{z}_n|\bar{J}_n \leq \bar{v}_n$ , where  $\bar{J}_1 = \{1, \dots, \text{ar}(S_1)\} \setminus J_1, \dots, \bar{J}_n = \{1, \dots, \text{ar}(S_n)\} \setminus J_n$ , we add as a disjunct of  $Q'$  the CQ

$$\exists \bar{y} S_{1,J_1,\bar{v}_1}(\bar{z}_1|J_1) \wedge \dots \wedge S_{n,J_n,\bar{v}_n}(\bar{z}_n|J_n).$$

It remains to prove that for all visible instances  $\mathcal{V}$ ,  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{NQL}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V})$ , and that the latter problem is active domain controllable.

For the easier direction, suppose that  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{false}$ , namely, that there is an  $\mathbf{S}$ -instance  $\mathcal{F}$  such that  $\mathcal{F} \models \mathcal{C}$ ,  $\mathcal{F} \models Q$ , and  $\text{Visible}(\mathcal{F}) = \mathcal{V}$ . We define the corresponding  $\mathbf{S}'$ -instance  $\mathcal{F}'$  as follows: for each relation  $R$  in  $\mathbf{S}$  and each copy  $R_{I,\bar{u}}$  of it in  $\mathbf{S}'$ , we instantiate  $R_{I,\bar{u}}$  with the set of tuples of the form  $\bar{a}|I$ , where  $\bar{a} \in R$ ,  $\bar{a}|\bar{I} \leq \bar{u}$ , and  $\bar{I} = \{1, \dots, \text{ar}(R)\} \setminus I$ . The instance  $\mathcal{F}'$  constructed in this way satisfies both the query  $Q'$  and the constraints in  $\mathcal{C}'$ . Moreover,  $\mathcal{F}'$  ranges over the active domain of  $\mathcal{V}$  and satisfies  $\text{Visible}(\mathcal{F}') = \text{Visible}(\mathcal{F}) = \mathcal{V}$ .

Conversely, suppose that  $\text{NQL}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V}) = \text{false}$ , namely, that there is an  $\mathbf{S}'$ -instance  $\mathcal{F}'$  such that  $\mathcal{F}' \models \mathcal{C}'$ ,  $\mathcal{F}' \models Q'$ , and  $\text{Visible}(\mathcal{F}') = \mathcal{V}$ . We need to give an instance for every relation  $R$  in  $\mathbf{S}$  so as to witness  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{false}$ . For the visible relations, we simply copy their content from the instance  $\mathcal{F}'$ . For the hidden relations  $R$ , the construction is more complicated, as it requires to merge the contents of the different copies  $R_{I,\bar{u}}$  in  $\mathbf{S}'$ . Formally, we fix an extension  $\mathbb{D}$  of the active domain of  $\mathcal{V}$  that contains  $k$  additional fresh values, where  $k$  is the maximum arity of the relations of  $\mathbf{S}$ . Then, we consider a fact  $R_{I,\bar{u}}(\bar{a})$  in the instance  $\mathcal{F}'$  and a candidate tuple  $\bar{b} \in \mathbb{D}^{\text{ar}(R)}$ . For each of such choices, we add the fact  $R(\bar{b})$  to  $\mathcal{F}$ , provided that  $I$  contains exactly those positions of  $\bar{b}$  that are associated with values in the active domain of  $\mathcal{V}$ ,  $\bar{b}|I = \bar{a}$ , and  $\bar{b}|\bar{I} \approx \bar{u}$ , where  $\bar{I} = \{1, \dots, \text{ar}(R)\} \setminus I$ . For example, if  $R$  has arity 5,  $I = \{1, 2\}$ ,  $\bar{u} = xyzczz$ , for some constant  $c \notin \text{adom}(\mathcal{V})$ , and  $R_{I,\bar{u}}$  contains the tuple  $\bar{a} = (a_1, a_2)$ , then we add to  $R$  all the tuples of the form  $\bar{b} = (a_1, a_2, a_3, c, a_3)$ , where  $a_3$  ranges over  $\mathbb{D} \setminus (\text{adom}(\mathcal{V}) \cup \{c\})$ . The instance  $\mathcal{F}$  constructed from  $\mathcal{F}'$  clearly agrees with  $\mathcal{F}'$  on the visible part.

Below we show that  $\mathcal{F}$  satisfies the constraints  $\mathcal{C}$  and the query  $Q$ . Consider any linear TGD of  $\mathcal{C}$  of the form

$$R(\bar{x}) \rightarrow \exists \bar{y} S(\bar{z})$$

and any fact  $R(\bar{a})$  that is the image under some homomorphism  $h$  of the left-hand side atom  $R(\bar{x})$ . Let  $I$  be the set of positions  $i \in \{1, \dots, \text{ar}(R)\}$  such that  $\bar{a}(i) \in \text{adom}(\mathcal{V})$  and let  $\bar{u} = \bar{a}|_{\bar{I}}$ , where  $\bar{I} = \{1, \dots, \text{ar}(R)\} \setminus I$ . By the previous constructions, we know that  $R_{I, \bar{u}}(\bar{a}|_I)$  is a fact of the instance  $\mathcal{F}'$ . Moreover,  $\mathcal{F}'$  verifies a linear TGD of the form

$$R_{I, \bar{u}}(\bar{x}|_I) \rightarrow \exists \bar{y} S_{J, \bar{v}}(\bar{z}|_J)$$

where  $J$  and  $\bar{v}$  can be chosen arbitrarily, provided that they satisfy the conditions 1) – 5) above. We derive the existence of a fact in  $\mathcal{F}'$  that is of the form  $S_{J, \bar{v}}(\bar{b})$ , where  $\bar{b}(j) = \bar{a}(i)$  whenever  $\bar{z}(j) = \bar{x}(i)$ . Again, by the previous constructions, we know that the tuple  $\bar{b}$  can be extended with values in  $\mathbb{D}$  so as to obtain a new tuple  $\bar{c}$  such that  $\bar{c} \leq \bar{v}$  that is the image of the right-hand side atom  $S(\bar{z})$  under some extension of the homomorphism  $h$ . This proves that  $\mathcal{F}$  satisfies the linear TGD  $R(\bar{x}) \rightarrow \exists \bar{y} S(\bar{z})$ .

Using similar arguments, one can show that  $\mathcal{F} \models Q$ , and hence  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{NQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V})$ . To conclude the proof, we observe that the active domain controllability of  $\text{NQI}(Q', \mathcal{C}', \mathbf{S}', \mathcal{V})$  follows from the two-way correspondence between the  $\mathbf{S}$ -instances  $\mathcal{F}$  and the  $\mathbf{S}'$ -instances  $\mathcal{F}'$  and from the fact that the latter instances  $\mathcal{F}'$  are directly constructed over the active domain of  $\mathcal{V}$ .  $\square$

Now we show how to exploit active domain controllability to prove that NQI problems can be solved not only efficiently, but “definably” using well-behaved query languages. For this, we introduce a variant of Datalog programs, called *GFP-Datalog* programs, whose semantics is given by greatest fixpoints. GFP-Datalog programs are defined syntactically in the same way as Datalog programs [AHV95], that is, as finite sets of rules of the form  $U(\bar{x}) \leftarrow Q(\bar{x})$  where the  $\bar{x}_i$  are implicitly universally quantified and  $Q$  is a conjunctive query whose free variables are exactly  $\bar{x}$ . As for Datalog programs, we distinguish between *extensional* (i.e., input) predicates and *intensional* (i.e., output) predicates. In the above rules we restrict the left-hand sides to contain only intensional predicates. Given a GFP-Datalog program  $P$ , the *immediate consequence operator* for  $P$  is the function that, given a database instance  $M$  consisting of both extensional and intensional relations, returns the database instance  $M'$  where the extensional relations are as in  $M$  and the tuples of each intensional relation  $U$  are those satisfying  $Q(M)$ , where  $Q$  is any query appearing on the right of a rule with  $U$ . The immediate consequence operator is monotone, and the semantics of the GFP-Datalog program on extensional database instance  $I$  is defined as the greatest fixpoint of this operator starting at the database instance  $I^+$  that extends  $I$  by setting each intensional relation “maximally” — that is, to the tuples of values from the active domain of  $I$  plus the constants appearing in the GFP-Datalog program. A program may also include a distinguished intensional predicate, the *goal predicate*  $G$ , and then the result is taken to be the projection of the greatest fixpoint onto  $G$ .



**Theorem 28.** *If  $Q$  is a UCQ,  $\mathcal{C}$  a set of linear TGDs (with constants), and  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  is active domain controllable, then  $\neg\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$ , viewed as a Boolean query over the visible part  $\mathcal{V}$ , is definable by a GFP-Datalog program that can be constructed in PTIME from  $Q$ ,  $\mathcal{C}$ , and  $\mathbf{S}$ .*

*Proof.* We need to describe by means of a GFP-Datalog program the function  $\neg\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  that maps an instance  $\mathcal{V}$  to either true or false depending on whether or not  $Q$  holds over some instance  $\mathcal{F}$  that satisfies the constraints  $\mathcal{C}$  and such that  $\text{Visible}(\mathcal{F}) = \mathcal{V}$ . Thanks to active domain controllability, it is sufficient consider only full instances constructed over the active domain of  $\mathcal{V}$ . More precisely, it is sufficient to show that a witnessing instance  $\mathcal{F}$  can be constructed as a greatest fixpoint starting from the values in the active domain of  $\mathcal{V}$ .

Below, we provide the GFP-Datalog program that computes  $\mathcal{F}$  starting from  $\mathcal{V}$ . The extensional relations are the ones in the visible part  $\mathcal{V}$ , while the intensional relations are the ones in the hidden part of the schema  $\mathbf{S}$ , plus an extra intensional relation  $A$  that collects the values in the active domain of  $\mathcal{V}$ . For each extensional relation  $R$  and each position  $i \in \{1, \dots, \text{ar}(R)\}$ , we add the rule  $A(x_i) \leftarrow R(\bar{x})$ , which derives all elements from the active domain into  $A$ . In addition, for each intensional relation  $R$ , we have the rule

$$R(\bar{x}) \leftarrow \bigwedge_i A(x_i) \wedge \bigwedge_{\substack{\text{linear TGD in } \mathcal{C} \text{ of the} \\ \text{form } R(\bar{x}) \rightarrow \exists \bar{y} S(\bar{z})}} S(\bar{z}) .$$

Let  $\mathcal{F}$  be the instance consisting of the visible part  $\mathcal{V}$  and the intensional relations  $R$  computed by the above Datalog program under the greatest fixpoint semantics. We claim that  $\mathcal{F}$  satisfies the constraints in  $\mathcal{C}$ . Indeed, if  $R(\bar{x}) \rightarrow \exists \bar{y} S(\bar{z})$  is a linear TGD in  $\mathcal{C}$  and  $R(\bar{a})$  is a fact of  $\mathcal{F}$ , with  $R(\bar{a})$  image of  $R(\bar{x})$  via some homomorphism  $h$ , then  $\mathcal{F}$  contains a fact of the form  $S(\bar{b})$ , where  $\bar{b}$  is the image of  $S(\bar{z})$  via some homomorphism  $h'$  that extends  $h$ .

To conclude, in order to compute the Boolean query  $\neg\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  starting from  $\mathcal{V}$ , we simply add to the above GFP-Datalog program one rule  $\text{Goal} \leftarrow S_1(\bar{z}_1) \wedge \dots \wedge S_n(\bar{z}_n)$  for each CQ  $\exists \bar{y} S_1(\bar{z}_1) \wedge \dots \wedge S_n(\bar{z}_n)$  of  $Q$ , and take  $\text{Goal}$  to be the final output of our program.  $\square$

Now, recall that the naïve fixpoint algorithm for a GFP-Datalog program takes exponential time in the maximum arity of the intensional relations, but only polynomial time in the size of the extensional relations and the number of rules. Thus, from Theorems 27 and 28, we immediately get:

**Corollary 29.** *If  $\mathcal{C}$  is restricted to range over sets of linear TGDs, then  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  has data complexity in PTIME and combined complexity in EXPTIME.*

**Example 4.** *Returning to the medical example from the introduction, Example 1, we see that the GFP-Datalog program is quite intuitive: since Patient is empty in the instance and we have a referential constraint from Appointment*

into **Patient**, **Appointment** is removed as well, leaving the empty instance. The program then simply evaluates the query on the resulting instance, which returns false, indicating that an NQI does hold on the original instance.

We do not know whether the use of GFP-Datalog can be replaced by other logics, such as Datalog. However we can show that in order to logically define  $\neg\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  from a given visible instance  $\mathcal{V}$ , it is necessary to go beyond first-order queries:

**Proposition 30.** *There are CQs  $Q$  and sets of IDs  $\mathcal{C}$  such that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  can not be described by a first-order query over  $\mathcal{V}$ . More generally, there are CQs  $Q$  and sets of IDs  $\mathcal{C}$  such that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{C}, \mathcal{V})$  is PTIME-hard in data complexity (that is, as  $\mathcal{V}$  varies over instances).*

*Proof.* To prove that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  is not first-order definable from  $\mathcal{V}$  we give a reduction from a graph reachability problem. The rough idea is to let some visible relations represent an input graph with two distinguished nodes playing the role of a source and a target. A proof for the existence of a path from the source to the target can be then exposed in the hidden relations. Formally, the nodes and the edges of the graph are encoded by two visible relations  $N$  and  $E$  of arity 1 and 2, respectively. The source and target nodes are encoded by two singleton relations  $A$  and  $B$ , respectively, of arity 1. A visible relation  $P$  of arity 5 is also introduced in order to drive the induction principle underlying the proof of existence of paths between pairs of nodes. This relation will contain the basic proof steps that can be used to witness reachability between two nodes. Formally,  $P$  contains tuples of the form  $(x, y, i, z, j)$ , where  $x, y, z$  are nodes in  $N$  and  $0 \leq i, j \leq |N|$ , such that:

1. either  $(y, z)$  is an edge and  $j = i + 1$ , meaning that if  $x$  is connected to  $y$  by a path of length  $i$ , then and  $(y, z)$  is an edge, then  $x$  is connected to  $z$  by a path of length  $j = i + 1$ ,
2. or  $x = y = z$  and  $i = j = 0$ , meaning that every node  $x$  is connected to itself by a path of length 0.

We fix  $\mathcal{V}$  to be our visible instance, which contains the relations  $N$ ,  $E$ , and  $P$ . In addition, we introduce a hidden relation  $T$  of arity 3, that will be constrained so as to contain only those triples  $(x, z, j)$  for which one can witness, using the basic proof steps in  $P$ , that  $x$  is connected to  $z$  by a path of length  $j$ . For this it suffices to enforce the following ID:

$$T(x, z, j) \rightarrow \exists y \ i \ T(x, y, i) \wedge P(x, y, i, z, j) .$$

It is easy to see that, for every full instance  $\mathcal{F}$  that satisfies the above constraint  $\mathcal{C}$  and agrees with  $\mathcal{V}$  in its visible part, if  $(x, z, j)$  is a tuple in  $T$ , then there is a path from  $x$  to  $z$  of length  $j$ . Conversely, if a node  $x$  is connected to a node  $y$  by a path of length  $j$ , then there is a way to extend the visible instance  $\mathcal{V}$  with a relation  $T$  that satisfies  $\mathcal{C}$  and contains the tuple  $(x, z, j)$ . Thus, if we let  $Q$  be the CQ  $\exists x \ z \ j \ A(x) \wedge B(z) \wedge T(x, z, j)$ , then we have  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$  iff the  $A$ -labelled node is connected to the  $B$ -labelled node. This property is clearly not definable in first-order logic.

A similar technique can be used to prove that  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{C}, \mathcal{V})$  is PTIME-hard for data complexity. The idea is to reduce the problem of evaluating a Boolean circuit to  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{C}, \mathcal{V})$ . The input and the structure of the Boolean circuit can be easily encoded in some visible relations. In addition, one introduces a visible relation  $P$  that contains all the valid rules that can be used during an evaluation. Finally, a hidden relation  $T$  can be used to expose a proof that the Boolean circuit evaluates to true.  $\square$

We give a tight EXPTIME lower bound for the combined complexity of NQL with linear TGDs:

**Theorem 31.** *The combined complexity of  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$ , where  $\mathcal{C}$  ranges over IDs, is EXPTIME-hard.*

*Proof.* We reduce the acceptance problem for an alternating PSPACE Turing machine  $M$  to  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$ . As in the proof of Theorem 7, we assume that the transition function of  $M$  maps each configuration to a set of exactly 2 successor configurations. In particular,  $M$  never halts. We also assume that  $M$  begins its computation with the head on the second position and never visits the first and last position of the tape. The acceptance condition of  $M$  is defined by distinguishing two special control states,  $q_{\text{acc}}$  and  $q_{\text{rej}}$ , that once reached will ‘freeze’  $M$  in its current configuration. We say that  $M$  accepts (the empty input) if for all paths in the computation tree, the state  $q_{\text{acc}}$  is eventually reached; otherwise, we say that  $M$  rejects.

Differently from the proofs of Theorem 7 and Theorem 9, the configurations of  $M$  can be described by simply specifying the label of each cell of the tape, the position of the head, and the control state of the Turing machine  $M$ . We thus define *cell values* as elements of  $V = (\Sigma \times Q) \uplus \Sigma$ , where  $\Sigma$  is the alphabet of  $M$  and  $Q$  is the set of its control states. If a cell has value  $(a, q)$ , this means that the associated letter is  $a$ , the control state of  $M$  is  $q$ , and the head is on this cell. Otherwise, if a cell has value  $a$ , this means that the associated letter is  $a$  and the head of  $M$  is not on this cell.

Now, let  $n$  be the size of the tape of  $M$ . We begin by describing the initial configuration of  $M$ . This is encoded by a visible relation  $C_0$  of arity  $n + 1$ , where the first attribute gives the identifier of the initial configuration and the remaining  $n$  attributes give the values of the tape cells. As the relation  $C_0$  is visible, we can immediately fix its content to be a singleton consisting of the tuple  $(x_0, y_1, y_2, y_3, \dots, y_n)$ , where  $x_0$  is the identifier of the initial configuration,  $y_1 = \perp$ ,  $y_2 = (\perp, q_0)$ ,  $y_3 = \dots = y_n = \perp$ . As for the other configurations of  $M$ , we store them into two distinct hidden relations  $C^\exists$  and  $C^\forall$ , depending on whether the control states are existential or universal. Each fact in one of these two relations consists of  $n + 1$  attributes, where the first attribute specifies an identifier and the remaining  $n$  attributes specify the cell values. We can immediately give the first constraint, which requires the initial configuration to be existential and stored also in the relation  $C^\exists$ :

$$C_0(x, y_1, \dots, y_n) \rightarrow C^\exists(x, y_1, \dots, y_n) .$$

To represent the computation tree of  $M$ , we encode pairs of subsequent configurations. In doing so, we not only store the identifiers of the configurations, but also their contents, in such a way that we can later check the correctness of the transitions using inclusion dependencies. We use different relations to recall the whether the current configuration is existential or universal and, in the latter case, whether the successor configuration is the first or the second one in the transition set (recall that the transition rules of  $M$  define exactly two successor configurations from each existential configuration). Formally, we introduce three hidden relations  $S^\exists$ ,  $S_1^\forall$ , and  $S_2^\forall$ , all of arity  $2n+2$ . We can easily enforce that the first  $n+1$  and the last  $n+1$  attributes in every tuple of  $S^\exists$ ,  $S_1^\forall$ , and  $S_2^\forall$  describe configurations in  $C^\exists$  and  $C^\forall$ :

$$\begin{aligned} S^\exists(x, \bar{y}, x', \bar{y}') &\rightarrow C^\exists(x, \bar{y}) & S^\exists(x, \bar{y}, x', \bar{y}') &\rightarrow C^\exists(x', \bar{y}') \\ S_1^\forall(x, \bar{y}, x', \bar{y}') &\rightarrow C^\forall(x, \bar{y}) & S_1^\forall(x, \bar{y}, x', \bar{y}') &\rightarrow C^\forall(x', \bar{y}') \\ S_2^\forall(x, \bar{y}, x', \bar{y}') &\rightarrow C^\forall(x, \bar{y}) & S_2^\forall(x, \bar{y}, x', \bar{y}') &\rightarrow C^\forall(x', \bar{y}') . \end{aligned}$$

Similarly, we guarantee that every existential (resp., universal) configuration has one (resp., two) successor configuration(s) in  $S^\exists$  (resp.,  $S_1^\forall$  and  $S_2^\forall$ ):

$$\begin{aligned} C^\exists(x, \bar{y}) &\rightarrow \exists x' \bar{y}' S^\exists(x, \bar{y}, x', \bar{y}') \\ C^\forall(x, \bar{y}) &\rightarrow \exists x' \bar{y}' S_1^\forall(x, \bar{y}, x', \bar{y}') \\ C^\forall(x, \bar{y}) &\rightarrow \exists x' \bar{y}' S_2^\forall(x, \bar{y}, x', \bar{y}') . \end{aligned}$$

We now turn to explaining how we can enforce the correctness of the transitions represented in the relations  $S^\exists$ ,  $S_1^\forall$ , and  $S_2^\forall$ . Compared to the proof of Theorem 7, the goal is simpler in this setting, as we can simply compare the values  $z_{-1}, z_0, z_{+1}$  for the cells at positions  $i-1, i, i+1$  in a configuration with the value  $z'$  for the cell at position  $i$  in the successor configuration. We thus introduce new visible relations  $N^\exists$ ,  $N_1^\forall$ , and  $N_2^\forall$  of arity 4. Each of these relations is initialized with the possible quadruples of cell values  $z_{-1}, z_0, z_{+1}, z'$  that are allowed by transition function of  $M$ . For example, if the transition function specifies that, when  $M$  is in the universal control state  $q$  and reads the letter  $a$ , then  $M$  spawns two subcomputations where the first one begins by rewriting  $a$  with  $a'$ , moving the head to the left, and switching to control state  $q'$ , then we add to  $N_1^\forall$  all the tuples of the form  $(a_{-1}, (a, q), a_{+1}, a')$  or  $(a_{-2}, a_{-1}, (a, q), (a_{-1}, q'))$ , with  $a_{-2}, a_{-1}, a_{+1} \in \Sigma$ . Accordingly, we introduce the following IDs, for all  $1 < i < n$ :

$$\begin{aligned} S^\exists(x, \bar{y}, x', \bar{y}') &\rightarrow N^\exists(y_{i-1}, y_i, y_{i+1}, y'_i) \\ S_1^\forall(x, \bar{y}, x', \bar{y}') &\rightarrow N_1^\forall(y_{i-1}, y_i, y_{i+1}, y'_i) \\ S_2^\forall(x, \bar{y}, x', \bar{y}') &\rightarrow N_2^\forall(y_{i-1}, y_i, y_{i+1}, y'_i) . \end{aligned}$$

Furthermore, we constrain the values of the extremal cells to never change:

$$\begin{aligned} S^\exists(x, \bar{y}, x', \bar{y}') &\rightarrow E(y_1, y'_1) & S^\exists(x, \bar{y}, x', \bar{y}') &\rightarrow E(y_n, y'_n) \\ S_1^\forall(x, \bar{y}, x', \bar{y}') &\rightarrow E(y_1, y'_1) & S_1^\forall(x, \bar{y}, x', \bar{y}') &\rightarrow E(y_n, y'_n) \\ S_2^\forall(x, \bar{y}, x', \bar{y}') &\rightarrow E(y_1, y'_1) & S_2^\forall(x, \bar{y}, x', \bar{y}') &\rightarrow E(y_n, y'_n) \end{aligned}$$

where  $E$  is another visible binary relation interpreted by the singleton instance  $\{(\perp, \perp)\}$ .

It remains to specify the query that checks that the Turing machine  $M$  reaches the rejecting state  $q_{\text{rej}}$  along some path of its computation tree. For this, we introduce a last visible relation  $V_{\text{rej}}$  that contains all cell values of the form  $(a, q_{\text{rej}})$ , with  $a \in \Sigma$ . The query that checks this property is

$$Q = \bigvee_{1 \leq i < n} \exists x \bar{y} \left( C^\exists(x, \bar{y}) \wedge V_{\text{rej}}(y_i) \right).$$

Let  $\mathcal{V}$  be the instance that captures the intended semantics of the visible relations  $V$ ,  $C_0$ ,  $N^\exists$ ,  $N_1^\forall$ ,  $N_2^\forall$ ,  $E$ , and  $V_{\text{rej}}$ . The proof that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$  iff  $M$  accepts (namely, has a computation tree where all paths visit the control state  $q_{\text{acc}}$ ) goes along the same lines of the proof of Theorem 7.  $\square$

## 4.2 Existence problems

Here we consider the complexity of the schema-level question,  $\exists \text{NQI}(Q, \mathcal{C}, \mathbf{S})$ . We first show that when the constraints are preserved under disjoint unions (e.g., connected frontier guarded TGDs), the existence of an NQI can be checked by considering a single “negative critical instance”, namely the empty visible instance  $\emptyset$ . This instance is easily seen to be realizable: the variant of the chase procedure that we introduced in Section 3.2 terminates immediately when initialized with the empty instance  $\mathcal{F}_0 = \emptyset$  and returns the singleton collection  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \emptyset)$  consisting of the empty  $\mathbf{S}$ -instance satisfying  $\mathcal{C}$ .

**Theorem 32.** *If the constraints  $\mathcal{C}$  are preserved under disjoint unions of instances, then  $\exists \text{NQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  iff  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \emptyset) = \text{true}$ .*

*Proof.* It is immediate to see that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \emptyset) = \text{true}$  implies  $\exists \text{NQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$ . We prove the converse implication by contraposition.

Suppose that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \emptyset) = \text{false}$ , namely, that there is an  $\mathbf{S}$ -instance  $\mathcal{F}$  satisfying  $\mathcal{C}$  and  $Q$  and such that  $\text{Visible}(\mathcal{F}) = \emptyset$ . We aim at proving that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{false}$  for all realizable visible instances  $\mathcal{V}$ . Let  $\mathcal{V}$  be such a realizable instance and let  $\mathcal{F}'$  be a  $\mathbf{S}$ -instance that satisfies  $\mathcal{C}$  and such that  $\text{Visible}(\mathcal{F}') = \mathcal{V}$ . We define the new instance  $\mathcal{F}''$  as a disjoint union of  $\mathcal{F}$  and  $\mathcal{F}'$ . Since the constraints  $\mathcal{C}$  are preserved under disjoint unions,  $\mathcal{F}''$  satisfies  $\mathcal{C}$ . Moreover,  $\mathcal{F}''$  satisfies the query  $Q$ , by monotonicity. Since  $\mathcal{V} = \text{Visible}(\mathcal{F}') = \text{Visible}(\mathcal{F}'')$ , we have  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{false}$ . Finally, since  $\mathcal{V}$  was chosen in an arbitrary way, this proves that  $\exists \text{NQI}(Q, \mathcal{C}, \mathbf{S}) = \text{false}$ .  $\square$

Using the “negative critical instance” result above and Theorem 22, we immediately see that  $\exists \text{NQI}(Q, \mathcal{C}, \mathbf{S})$  is decidable in 2EXPTIME for GNFO constraints that are closed under disjoint unions, and in particular for connected frontier-guarded TGDs. Combining with Corollary 29 also gives an EXPTIME bound for linear TGDs. In fact, we can improve this upper by observing that the NQI problem over the empty visible instance reduces to classical Open-World Query answering:

**Proposition 33.** *For any Boolean CQ  $Q$ ,  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \emptyset)$  holds iff  $\text{OWQ}(Q', \mathcal{C}, \text{CanonDB}(Q))$  holds, where*

$$Q' = \bigvee_{R \in \mathbf{S}_v} \exists \bar{x} R(\bar{x})$$

and  $\text{CanonDB}(Q)$  is the canonical database of the CQ  $Q$ .

*Proof.* Suppose that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \emptyset) = \text{true}$ . This means that every  $\mathbf{S}$ -instance that satisfies the constraints in  $\mathcal{C}$  and has empty visible part, must violate the query  $Q$ . By contraposition, every  $\mathbf{S}$ -instance that satisfies the constraints  $\mathcal{C}$  and contains  $\text{CanonDB}(Q)$  (i.e., satisfies  $Q$ ), must contain some visible facts, and hence satisfy the UCQ  $Q'$ . This implies that  $\text{OWQ}(Q', \mathcal{C}, \text{CanonDB}(Q)) = \text{true}$ .

The proof that  $\text{OWQ}(Q', \mathcal{C}, \text{CanonDB}(Q)) = \text{true}$  implies  $\exists \text{NQI}(Q, \mathcal{C}, \mathbf{S}, \emptyset) = \text{true}$  follows symmetric arguments.  $\square$

We know from previous results [BGO10] that OWQ for Boolean UCQs and linear TGDs is in PSPACE. From the above reduction, we immediately get that the problem  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \emptyset)$ , and hence (by Theorem 32) the problem  $\exists \text{NQI}(Q, \mathcal{C}, \mathbf{S})$ , for a set of linear TGDs is also in PSPACE.

**Corollary 34.** *The problem  $\exists \text{NQI}(Q, \mathcal{C}, \mathbf{S})$ , as  $Q$  ranges over Boolean UCQ and  $\mathcal{C}$  over sets of linear TGDs, is in PSPACE.*

Matching lower bounds for  $\exists \text{NQI}$  come by a converse reduction from Open-World Query answering.

To prove this reduction, we first provide a characterization of the  $\text{NQI}$  problem over the empty visible instance, which is based, like Proposition 12, on our chase procedure:

**Proposition 35.** *If  $Q$  is a Boolean CQ and  $\mathcal{C}$  is a set of TGDs and EGDs without constants over a schema  $\mathbf{S}$ , then  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \emptyset) = \text{true}$  iff either  $Q$  contains a visible atom, or it does not and in this case  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \text{CanonDB}(Q)) = \emptyset$ .*

*Proof.* We give first the proof when  $\mathcal{C}$  consists only of TGDs. Suppose that  $Q$  does not contain visible atoms and  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \text{CanonDB}(Q))$  contains an instance  $K$ . Because every instance in  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \text{CanonDB}(Q))$  satisfies the constraints in  $\mathcal{C}$  and the query  $Q$ , and it has an empty visible part by Lemma 11, we conclude that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \emptyset) = \text{false}$ . Conversely, suppose that  $\text{NQI}(Q, \mathcal{C}, \mathbf{S}, \emptyset) = \text{false}$ . This means that there is an  $\mathbf{S}$ -instance  $\mathcal{F}$  with no visible facts that satisfies the constraints in  $\mathcal{C}$  and the query  $Q$ . Since  $\mathcal{F} \models Q$ , there is a homomorphism  $g$  from  $\text{CanonDB}(Q)$  to  $\mathcal{F}$ . Moreover, since  $Q$  contains no visible atoms, the two instances  $\mathcal{F}$  and  $\text{CanonDB}(Q)$  agree on the visible part. By Lemma 11, letting  $\mathcal{F}_0 = \text{CanonDB}(Q)$ , we get the existence of an instance  $K$  in  $\text{Chases}_{\text{vis}}(\mathcal{C}, \mathbf{S}, \text{CanonDB}(Q))$ .

In the presence of EGDs, we apply the extension of Lemma 11 to TGDs and EGDs, as discussed earlier.  $\square$

As in the positive case, the upper bounds are tight:

**Theorem 36.**  $\exists\text{NQL}(Q, \mathcal{C}, \mathbf{S})$  is 2EXPTIME-hard as  $Q$  ranges over Boolean CQs and  $\mathcal{C}$  over sets of connected FGTGDs.

**Theorem 37.**  $\exists\text{NQL}(Q, \mathcal{C}, \mathbf{S})$  is PSPACE-hard as  $Q$  ranges over Boolean CQs and  $\mathcal{C}$  over sets of linear TGDs.

The first theorem is proven by reducing the open-world query answering problem to  $\exists\text{NQL}$ , and then applying a prior 2EXPTIME-hardness result from Cali et al. [CGK13]. The PSPACE lower bound is shown by a reduction from the implication problem for IDs, shown PSPACE-hard by Casanova et al. [CFP84].

To prove lower bounds for  $\exists\text{NQL}$ , we first give a reduction from Open-World Query answering:

**Proposition 38.** *There is a polynomial time reduction from the Open-World Query answering problem over a set of connected FGTGDs without constants and a connected Boolean CQ to an  $\exists\text{NQL}$  problem over a set of connected FGTGDs without constants and a Boolean CQ.*

*Proof.* Consider the Open-World Query answering problem over a schema  $\mathbf{S}$ , a set  $\mathcal{C}$  of constraints without constants and closed under disjoint union, a Boolean CQ  $Q$ , and a  $\mathbf{S}$ -instance  $\mathcal{F}$ . We reduce this problem to an  $\exists\text{NQL}$  problem over a new schema  $\mathbf{S}'$ , a new set of constraints  $\mathcal{C}'$ , and a new Boolean CQ  $Q'$ . The schema  $\mathbf{S}'$  is obtained from  $\mathbf{S}$  by adding a relation **Good** of arity 0, which is assumed to be the only visible relation in  $\mathbf{S}'$ . The set of constraints  $\mathcal{C}'$  is equal to  $\mathcal{C}$  unioned with the constraint

$$S_1(\bar{x}_1) \wedge \dots \wedge S_m(\bar{x}_m) \rightarrow \text{Good}$$

where  $S_1(\bar{x}_1), \dots, S_m(\bar{x}_m)$  are the atoms in the CQ  $Q$ . The query  $Q'$  is defined as the *canonical query* of the instance  $\mathcal{F}$ , obtained by replacing each value  $v$  with a variable  $y_v$  and by quantifying existentially over all these variables. Note that  $\text{CanonDB}(Q')$  is isomorphic to the input instance  $\mathcal{F}$ .

Now, assume that the original constraints in  $\mathcal{C}$  were connected FGTGDs and the CQ  $Q$  was also connected. By construction, the constraints in  $\mathcal{C}'$  turn out to be also connected FGTGDs. In particular, the satisfiability of these constraints is preserved under disjoint unions, and hence from Theorem 32,  $\exists\text{NQL}(Q', \mathcal{C}', \mathbf{S}') = \text{true}$  iff  $\text{NQL}(Q', \mathcal{C}', \mathbf{S}', \emptyset) = \text{true}$ . Thus, it remains to show that  $\text{NQL}(Q', \mathcal{C}', \mathbf{S}', \emptyset) = \text{true}$  iff  $\text{OWQ}(Q, \mathcal{C}, \mathcal{F}) = \text{true}$ .

By contraposition, suppose that  $\text{OWQ}(Q, \mathcal{C}, \mathcal{F}) = \text{false}$ . This means that there is a  $\mathbf{S}$ -instance  $\mathcal{F}'$  that contains  $\mathcal{F}$ , satisfies the constraints in  $\mathcal{C}$ , and violates the query  $Q$ . In particular,  $\mathcal{F}'$ , seen as an instance of the new schema  $\mathbf{S}'$ , without the visible fact **Good**, satisfies the query  $Q'$  and the constraints in  $\mathcal{C}'$  (including the constraint that derives **Good** from the satisfiability of  $Q$ ). The  $\mathbf{S}'$ -instance  $\mathcal{F}'$  thus witnesses the fact that  $\text{NQL}(Q', \mathcal{C}', \mathbf{S}', \emptyset) = \text{false}$ .

Conversely, suppose that  $\text{NQL}(Q', \mathcal{C}', \mathbf{S}', \emptyset) = \text{false}$ . Recall that the constraints in  $\mathcal{C}'$  do not use constants and  $Q'$  contains no visible facts. We can thus apply Proposition 35 and derive  $\text{Chases}_{\text{vis}}(\mathcal{C}', \mathbf{S}', \text{CanonDB}(Q')) \neq \emptyset$ . Note

that  $\text{CanonDB}(Q')$  is clearly isomorphic to the original instance  $\mathcal{F}$ . In particular, there is an instance  $K$  in  $\text{Chases}_{\text{vis}}(\mathcal{C}', \mathbf{S}', \text{CanonDB}(Q'))$  that contains the original instance  $\mathcal{F}$ , satisfies the constraints in  $\mathcal{C}'$ , and does not contain the visible fact **Good**. From the latter property, we derive that  $K$  violates the query  $Q$ . Thus  $K$ , seen as an instance of the schema  $\mathbf{S}$ , witnesses the fact that  $\text{OWQ}(Q, \mathcal{C}, \mathcal{F}) = \text{false}$ .  $\square$

We note that there are two variants of OWQ, corresponding to finite and infinite instances. However, by finite-controllability of FGTGDs, inherited from the finite model property of GNFO (see Theorem 1) these two variants agree. Hence we do not distinguish them. Similar remarks hold for other uses of OWQ within proofs in the paper.

We are now ready to prove Theorem 36, namely, the 2EXPTIME-hardness of the problem  $\exists\text{NQL}(Q, \mathcal{C}, \mathbf{S})$ , where  $Q$  ranges over Boolean CQs and  $\mathcal{C}$  ranges over sets of connected FGTGDs.

*Proof of Theorem 36.* Theorem 6.2 of Calì et al. [CGK13] shows 2EXPTIME-hardness of open-world query answering for FGTGDs. An inspection of the proof shows that only connected FGTGDs are required. Thus, the theorem follows immediately from Proposition 38.  $\square$

We now turn towards proving Theorem 37, namely, the PSPACE lower bound for  $\exists\text{NQL}$  under linear TGDs. Recall that the reduction in Proposition 38 does not preserve smaller constraint classes, such as linear TGDs. We thus prove the theorem using a separate reduction.

*Proof of Theorem 37.* We reduce from the implication problem for inclusion dependencies, which is known to be PSPACE-hard from Casanova et al. [CFP84]. Consider a set of IDs  $\mathcal{C}$  and an additional ID  $\delta = S_*(x_*) \rightarrow \exists \bar{y} T_*(\bar{z}_*)$ , where  $\bar{x}_*, \bar{y}$  are sequences of pairwise distinct variables and  $\bar{z}_*$  is a sequence of variables from  $\bar{x}_*$  and  $\bar{y}$ . Note that we annotated with the subscript  $*$  the relations and variables in  $\delta$  in order to make it clear when refer later to these particular objects.

We create a new schema  $\mathbf{S}'$  that contains, for each relation  $R$  of arity  $k$  in the original schema  $\mathbf{S}$ , a relation  $R'$  of arity  $2k$ . We also add to  $\mathbf{S}'$  a copy of each relation  $R$  in  $\mathbf{S}$ , without changing the arity. Furthermore, we add a 0-ary relation **Good**, which is the only visible relation of  $\mathbf{S}'$ . For each ID in  $\mathcal{C}$  of the form

$$R(\bar{x}) \rightarrow \exists \bar{y} S(\bar{z})$$

we introduce a corresponding ID in  $\mathcal{C}'$  of the form

$$R'(\bar{x}, \bar{x}') \rightarrow \exists \bar{y} S'(\bar{z}, \bar{x}')$$

where the variables in  $\bar{x}'$  are distinct from the variables in  $\bar{x}$ . We also add the constraints

$$\begin{aligned} R(\bar{x}) &\rightarrow R'(\bar{x}, \bar{x}) \\ T'_*(\bar{z}_*, \bar{x}') &\rightarrow \text{Good} \end{aligned}$$



where the elements of  $\bar{z}_\star$  are arranged as in the atom  $T_\star(\bar{z}_\star)$  that appears on the right-hand side of the ID  $\delta$ . Note that the constraint that copies the content from  $R$  to  $R'$  and duplicates the attributes is not an ID, but is still a linear TGD. The query of our  $\exists\text{NQI}$  problem is defined as

$$Q' = \exists \bar{x} S_\star(\bar{x}) .$$

Note that the constraints that we just defined are preserved under disjoint unions. Thus, by Theorem 32, we know that  $\exists\text{NQI}(Q', \mathcal{C}', \mathbf{S}') = \text{true}$  iff  $\text{NQI}(Q', \mathcal{C}', \mathbf{S}', \emptyset) = \text{true}$ . Below, we prove that the latter holds iff the ID  $\delta$  is implied by the set of IDs in  $\mathcal{C}$ .

In one direction, suppose that the implication holds. From this, we can easily infer in the schema  $\mathbf{S}'$  the following dependency:

$$S'_\star(\bar{x}, \bar{x}) \rightarrow \exists \bar{y} T'_\star(\bar{z}_\star, \bar{x})$$

Consider now a full  $\mathbf{S}'$ -instance  $\mathcal{F}'$  with empty visible part. We show that the query  $Q'$  is not satisfied, namely,  $\mathcal{F}'$  cannot contain a fact of the form  $S_\star(\bar{x})$ . If it did, then, by the copy of the constraints on the primed relations, this would yield the fact  $S'_\star(\bar{x}_\star, \bar{x}_\star)$ , and hence, by the constraints, also the facts  $T'_\star(\bar{z}_\star, \bar{x}_\star)$  and **Good**. This however would contradict the hypothesis that  $\mathcal{F}'$  has empty visible part.

In the other direction, suppose that the implication fails and consider a witness  $\mathbf{S}$ -instance  $\mathcal{F}$  that contains the fact  $S_\star(\bar{x}_\star)$  but not the corresponding  $T_\star$  fact. We create a full  $\mathbf{S}'$ -instance  $\mathcal{F}'$  with empty visible part where  $Q'$  holds, thus showing that  $\exists\text{NQI}(Q', \mathcal{C}', \mathbf{S}', \emptyset) = \text{false}$ . We first copy in  $\mathcal{F}'$  the content of all relations  $R$  from  $\mathcal{F}$ . In particular,  $\mathcal{F}'$  contains the fact  $S_\star(\bar{x}_\star)$ , but no  $T_\star$  fact. The primed relations  $R'$  in  $\mathcal{F}'$  are set to contain all and only the facts of the form  $R'(\bar{x}, \bar{x}_\star)$ , where  $R(\bar{x})$  is a fact in  $\mathcal{F}$ . Finally, we set **Good** to be the empty relation in  $\mathcal{F}'$ . Clearly,  $Q$  holds in  $\mathcal{F}'$  and the visible part is the empty instance. It is also easy to verify that all the constraints in  $\mathcal{C}'$  are satisfied by  $\mathcal{F}'$ , and this completes the proof.  $\square$

Note that the reduction above does not create a schema with IDs, but rather with general linear TGDs (variables can be repeated on the right). We do not know whether  $\exists\text{NQI}(Q, \mathcal{C}, \mathbf{S})$  is PSPACE-hard even for constraints consisting of IDs.

We can easily see that the connectedness requirement is critical for decidability:

**Theorem 39.** *The problem  $\exists\text{NQI}(Q, \mathcal{C}, \mathbf{S})$  is undecidable as  $Q$  ranges over Boolean CQs and  $\mathcal{C}$  over sets of FGTGDs.*

*Proof.* We give a reduction from the *model conservativity problem* for  $\mathcal{EL}$  TBoxes, which is shown undecidable in [LW07]. Intuitively,  $\mathcal{EL}$  is a logic that defines FGTGDs over relations of arity 2, called “TBoxes”. Given some TBoxes  $\phi_1$  and  $\phi_2$  over two schemas  $\mathbf{S}_1$  and  $\mathbf{S}_2$ , respectively, with  $\mathbf{S}_1 \subseteq \mathbf{S}_2$ , we say that  $\phi_2$  is a *model conservative extension* of  $\phi_1$  if every  $\mathbf{S}_1$ -instance  $\mathcal{V}$  that satisfies

$\phi_1$  can be extended to an  $\mathbf{S}_2$ -instance that satisfies  $\phi_2$  without changing the interpretation of the predicates in  $\mathbf{S}_1$ , that is, by only adding an interpretation for the relations that are in  $\mathbf{S}_2$  but not in  $\mathbf{S}_1$ . The model conservativity problem consists of deciding whether  $\phi_2$  is a model conservative extension of  $\phi_1$ . The proof in [LW07] shows that this problem is undecidable for both finite instances and arbitrary instances.

We reduce the above problem to the complement of  $\exists\text{NQI}(Q, \mathcal{C}, \mathbf{S})$ , for suitable  $Q$ ,  $\mathcal{C}$ , and  $\mathbf{S}$ , as follows. Given some TBoxes  $\phi_1$  and  $\phi_2$  over the schemas  $\mathbf{S}_1 \subseteq \mathbf{S}_2$ , let  $\mathbf{S}$  be the schema obtained from  $\mathbf{S}_2$  by adding a new predicate **Good** of arity 0 and by letting the visible part be  $\mathbf{S}_1$  (in particular, the relation **Good** is hidden). Further let  $\mathcal{C} = \{\phi_1, \text{Good} \rightarrow \phi_2\}$ , where  $\text{Good} \rightarrow \phi_2$  is shorthand for the collection of FGTGDs obtained by adding **Good** as a conjunct to the left-hand side of each constraint of  $\phi_2$  (note that this makes the constraints unconnected). Finally, consider the query  $Q = \text{Good}$ . We have that  $\exists\text{NQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  iff there is an  $\mathbf{S}_1$ -instance  $\mathcal{V}$  satisfying  $\phi_1$ , none of whose  $\mathbf{S}_2$ -expansions satisfies  $\phi_2$ .  $\square$

### 4.3 Summary for Negative Query Implication

A summary of results on negative implication is below. We notice that the decidable cases are orthogonal to those for positive implications. Note also that unlike in the positive cases, we have tractable cases for data complexity.

	NQI Data	NQI Combined	$\exists\text{NQI}$
Linear TGD	PTime-cmp Cor 29/Prop 30	EXPTIME-cmp Cor 29/Thm 31	PSpace-cmp Cor. 34/Thm 37
Conn. Disj. FGTGD	EXPTIME-cmp Thm 22/Thm 23	2EXPTIME-cmp Thm 22/Thm 23	2EXPTIME-cmp Thm 32/Thm 36
FGTGD & GNFO	EXPTIME-cmp Thm 22/Thm 23	2EXPTIME-cmp Thm 22/Thm 23	undecidable Thm 39

## 5 Extensions and special cases

We present some results concerning natural extensions of the framework.

First note, that throughout this work we have restricted to queries given by Boolean UCQs. The natural extension of the notion of query implication for non-Boolean queries is to consider disclosure of information concerning membership of any visible tuple in the query output. E.g.  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  would hold if for some tuple  $\bar{t}$  in the active domain of  $\mathcal{V}$ ,  $\bar{t} \in Q(\mathcal{F})$  for any instance  $\mathcal{F}$  of  $\mathbf{S}$  satisfying the constraints and having visible part  $\mathcal{V}$ . We show that *all of our results carry over to the non-boolean case*.

The natural extension of the notion of query implication for non-Boolean queries is to consider disclosure of information concerning membership of any visible tuple in the query output. E.g.  $\text{PQI}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  would hold if for some tuple  $\bar{t}$  in the active domain of  $\mathcal{V}$ ,  $\bar{t} \in Q(\mathcal{F})$  for any instance  $\mathcal{F}$  of  $\mathbf{S}$  satisfying the constraints and having visible part  $\mathcal{V}$ . Since the lower-bounds for Boolean

problems are clearly inherited by the non-Boolean ones, we focus on whether the upper bounds carry over.

All the complexity upper bounds for the instance-level problem carry over straightforwardly using the simple approach of substituting in each potential output a tuple from  $\mathcal{V}$  and utilizing the prior algorithms on the resulting Boolean queries. The complexity for each substitution preserves the upper bounds since they hold in the presence of constants, and the iteration over tuples can be absorbed in the complexity classes given in our upper bounds: for data complexity the iteration is polynomial, while for combined complexity the number of tuples can be exponential, but our bounds are at least exponential. Further, GFP-Datalog definability for negative implications also extends straightforwardly to the non-Boolean case: Theorem 27 extends with the same statement and proof, while the argument in Theorem 28 is easily extended to show that there is a GFP-Datalog program that returns the complement of  $\text{NQL}(Q, \mathcal{C}, \mathbf{S})$  within the active domain.

The complexity results for  $\exists\text{PQL}$  also generalize to the non-Boolean case: we can revise Theorem 10 to state  $\exists\text{PQL}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  iff there is a positive query implication for the tuple  $(a, \dots, a)$  and the instance  $\mathcal{V}_{\{a\}}$ . For  $\exists\text{NQL}$ , we can extend Theorem 32 to show that for constraints preserved under disjoint union, if there is a positive query implication involving some visible instance  $\mathcal{V}$  and a tuple  $\bar{t}$ , then there is one involving the empty instance and some tuple  $\bar{t}$ . From this it follows that the complexity bounds for  $\exists\text{NQL}$  carry over to the non-Boolean case.

**Beyond conjunctive queries.** So far we have considered only the case where  $Q$  is a UCQ. It is natural to extend the query language even further, to Boolean combinations of Boolean conjunctive queries (BCCQs). We note that the problem  $\text{PQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$ , as  $Q$  ranges over BCCQs, subsumes both  $\text{PQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  and  $\text{NQL}(Q, \mathcal{C}, \mathbf{S}, \mathcal{V})$  for  $Q$  a UCQ. Thus all lower bounds for either of these two problems are inherited by the BCCQ problem. The corresponding instance level problems are still decidable. Indeed, this holds even when  $Q$  is a GNFO sentence – we can just apply using the same translation to GNFO satisfiability applied in Theorems 4 and 22. However, for the schema-level problems  $\exists\text{PQL}$  and  $\exists\text{NQL}$  we immediately run into problems:

**Theorem 40.** *The problem  $\exists\text{PQL}(Q, \mathcal{C}, \mathbf{S})$  for a Boolean combination  $Q$  of CQs is undecidable, even when the constraints are IDs. The same holds for  $\exists\text{NQL}(Q, \mathcal{C}, \mathbf{S})$ .*

*Proof.* As in the previous undecidability results, we reduce a tiling problem with tiles  $T$ , initial tile  $t_0 \in T$  and horizontal and vertical constraints  $H, V \subseteq T \times T$  to the problem  $\exists\text{PQL}(Q, \mathcal{C}, \mathbf{S})$ . Again, for convenience we deal with the infinite variant of the problem. The idea will be that the visible instance witnessing  $\exists\text{PQL}$  represents the tiling, and invisible instances represent challenges to the correctness of the tiling.

We model the infinite grid to be tiled by visible relations  $E_H$  and  $E_V$ , and the tiling function by a collection of unary visible relations  $U_t$ , for all tiles  $t \in T$ .

The invisible relations represent markings of the grid for possible errors. There are several kinds of challenges. We focus on the horizontal consistency challenge, which selects two nodes in the  $E_H$  relation, to challenge whether the nodes satisfy the horizontal constraint. Formally, the challenge is captured by a binary invisible predicate  $\text{HorChallenge}(x, y)$ , with an associated integrity constraint

$$\text{HorChallenge}(x, y) \rightarrow E_H(x, y) .$$

The query  $Q$  will be satisfied only when the following *negated* CQs hold, for all pairs  $(t, t') \notin H$ :

$$\neg \exists x y \text{ HorChallenge}(x, y) \wedge U_t(x) \wedge U_{t'}(y) .$$

Note that this can only happen if the relation  $\text{HorChallenge}$  has selected two horizontally adjacent nodes whose tiles violate the horizontal constraints. The vertical constraints are enforced in a similar way using an invisible relation  $\text{VertChallenge}$  and another negated CQ.

Recall that in the infinite grid, we have unique vertical and horizontal successors of each node, and the horizontal and vertical successor functions commute. Thus far we have not enforced that  $E_V$  and  $E_H$  have this property. We will use additional hidden relations and IDs to enforce that every element is related to at least one other via  $E_H$  and  $E_V$ .

We first show how to enforce that every element has at most one horizontal successor (“functionality challenge”). We introduce a hidden relation  $\text{HorFuncChallenge}(x, y, y')$  and constraints

$$\begin{aligned} \text{HorFuncChallenge}(x, y, y') &\rightarrow E_H(x, y) \\ \text{HorFuncChallenge}(x, y, y') &\rightarrow E_V(x, y') . \end{aligned}$$

We also add to the query  $Q$  the conjunct:

$$\left( \neg \exists x y y' \text{ HorFuncChallenge}(x, y, y') \right) \vee \left( \exists x y \text{ HorFuncChallenge}(x, y, y) \right) .$$

We claim that if there is a visible instance witnessing  $\exists \text{PQI}$ , then  $E_H$  is functional. Indeed, if  $E_H$  were not functional in the visible instance, then we could choose a node  $x$  with two distinct  $E_H$ -successors  $y$  and  $y'$ , add only the tuple  $(x, y, y')$  to  $\text{HorFuncChallenge}$ , and obtain a full instance that satisfies the constraints but not the query  $Q$ . Conversely, suppose that  $E_H$  is functional in a visible instance  $\mathcal{V}$ , and consider any full instance  $\mathcal{F}$  that satisfies the constraints and agrees with  $\mathcal{V}$  on the visible part. If there are no tuples in  $\text{HorFuncChallenge}$ , the conjunct above is clearly satisfied by its first disjunct. If there is some tuple  $(x, y, y')$  in  $\text{HorFuncChallenge}$ , then by the constraints, we must have  $E_H(x, y)$  and  $E_H(x, y')$ , and hence, by functionality,  $y = y'$ . In this case, the conjunct above holds via the second disjunct. The functionality of the vertical relation  $E_V$  is enforced in an analogous way.

Commutativity of  $E_H$  and  $E_V$  can be also enforced using a similar technique. We add a hidden relation  $\text{ConfChallenge}(x, y, z, u, v)$  with constraints:

$$\begin{aligned}\text{ConfChallenge}(x, y, z, u, v) &\rightarrow E_H(x, y) \\ \text{ConfChallenge}(x, y, z, u, v) &\rightarrow E_V(y, u) \\ \text{ConfChallenge}(x, y, z, u, v) &\rightarrow E_V(x, z) \\ \text{ConfChallenge}(x, y, z, u, v) &\rightarrow E_H(z, v) .\end{aligned}$$

A potential tuple in  $\text{ConfChallenge}(x, y, z, u, v)$  represents the join of a triple of nodes moving first horizontally and then vertically from  $x$  (i.e.,  $x, y, u$ ) and a triple going first vertically and then horizontally from  $x$  (i.e.,  $x, z, v$ ). For the relations to commute, we must satisfy the query

$$\left( \neg \exists x y z u v \text{ConfChallenge}(x, y, z, u, v) \right) \vee \left( \exists x y z u \text{ConfChallenge}(x, y, z, u, u) \right)$$

in the full instance. Thus, we add the above conjunct to  $Q$ .

Putting the various components of  $Q$  for different challenges together as a Boolean combination of CQ, completes the proof of the theorem.  $\square$

**The case of conjunctive query views.** As mentioned earlier, the database community has studied the PQI problem in the case where the constraints consist exactly of CQ-view definitions defining each visible relation in terms of invisible relations. Formally, a CQ-view based scenario consists of a schema  $\mathbf{S} = \mathbf{S}_v \cup \mathbf{S}_h$ , namely, the union of a schema for the visible relations and a schema for the hidden relations, and a set of constraints  $\mathcal{C}$  between visible and hidden relations that must be of a particular form. For each visible relation  $R \in \mathbf{S}_v$ ,  $\mathcal{C}$  must contain two dependencies of the form

$$\begin{aligned}R(\bar{x}) &\rightarrow \exists \bar{y} \phi_R(\bar{x}, \bar{y}) \\ \phi_R(\bar{x}, \bar{y}) &\rightarrow R(\bar{x})\end{aligned}$$

where  $\phi_R$  is a conjunction of atoms over the hidden schema  $\mathbf{S}_h$ . Furthermore, all constraints in  $\mathcal{C}$  must be of the above forms. Note that this CQ-view scenario is incomparable in expressiveness to GNFO constraints.

The instance-level problems are still well-behaved, because given a visible instance  $\mathcal{V}$ , the constraints can be rewritten as  $\mathcal{C}_1 \wedge \mathcal{C}_2$ , where  $\mathcal{C}_1$  consists of TGDS from the view relations to the base relations, and  $\mathcal{C}_2$  consists of disjunctive linear EGDs from the base relations to the various possible tuples in the view relations. Thus the “disjunctive chase” of  $\mathcal{V}$  with these constraints will terminate after a finite number of rounds. From this we can directly argue that a counterexample superinstance for either PQI and NQI must be of polynomial size.

The decidability of the  $\exists$ PQI problem follows immediately from these observations and Theorem 10, which applies to constraints capturing CQ-view definitions. In contrast, for the  $\exists$ NQI problem we prove that

**Theorem 41.** *The  $\exists$ NQI problem under constraints given as CQ-view definitions is undecidable.*

*Proof.* We given a reduction from a tiling problem that is specified by a set of tiles  $T$ , an initial tile  $t_\perp \in T$ , and horizontal and vertical constraints  $H, V \subseteq T \times T$ . As before, we will deal for simplicity with the infinite variant, thus considering the problem of tiling the infinite grid  $\mathbb{N} \times \mathbb{N}$ .

As usual, we will have visible relations  $E_H$  and  $E_V$  representing the horizontal and vertical edges of the grid. As these relations must be associated with CQ-view definitions, we add hidden copies  $E'_H$  and  $E'_V$  of them and enforce the trivial dependencies:

$$\begin{aligned} E_H(x, y) &\leftrightarrow E'_H(x, y) \\ E_V(x, y) &\leftrightarrow E'_V(x, y) . \end{aligned}$$

Similarly, each node of the grid has to be associated with a tile in  $T$ , and this will be represented by a tuple of visible unary relations  $U_t$  and hidden copies  $U'_t$ , and constrain them with the dependencies  $U_t(x) \leftrightarrow U'(t)$ , for all  $t \in T$ .

As in earlier undecidability results, such as Theorem 40, the first goal is to ensure that for each node, there exists at most one predecessor and at most one successor for the relations  $E_H$  and  $E_V$ . We explain how to ensure this for the successor case and the relation  $E_H$ , but similar constructions work for the other cases. We introduce a hidden relation **HorFuncChallenge** of arity 4, and a visible relation **ErrHorFun** of arity 3 with the associated CQ-view definition

$$\text{ErrHorFun}(x, y, x') \leftrightarrow \text{HorFuncChallenge}(x, y, x', y) .$$

Our query  $Q$  will contain as a conjunct the following UCQ:

$$\begin{aligned} Q_{\text{HorFuncChallenge}} = & \left( \exists x y y' \text{ErrHorFun}(x, y, y') \right) \vee \\ & \left( \exists x y y' \text{HorFuncChallenge}(x, y, x', y) \wedge E_H(x, y) \wedge E_H(x, y') \right) . \end{aligned}$$

We explain how the subquery  $Q_{\text{HorFuncChallenge}}$  enforces that every element has at most one successor in the relation  $E_H$ .

Suppose that  $\exists \text{NQL}(Q_{\text{HorFuncChallenge}}, \mathcal{C}, \mathbf{S}) = \text{true}$ , namely, that there exists an  $\mathbf{S}_v$ -instance  $\mathcal{V}$  such that  $\text{NQL}(Q_{\text{HorFuncChallenge}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ . The visible relation **ErrHorFun** must be empty in  $\mathcal{V}$ , as otherwise the query  $Q_{\text{HorFuncChallenge}}$  would be satisfied in every full instance that agrees with  $\mathcal{V}$  on the visible part (note that  $\mathcal{V}$  is clearly realizable). Moreover, as **ErrHorFun** is empty in  $\mathcal{V}$ , every full instance that satisfies the constraints and agrees with  $\mathcal{V}$  does not contain a fact of the form **HorFuncChallenge**( $x, y, x', y$ ). Now, suppose, by way of contradiction, that there is an element  $x$  with two distinct  $E_H$ -successors  $y$  and  $y'$ . We can construct a full instance that extends  $\mathcal{V}$  with the single fact **HorFuncChallenge**( $x, y, x, y'$ ). This full instance satisfies all the constraints in  $\mathcal{C}$  and also the query  $Q_{\text{HorFuncChallenge}}$ , thus contradicting  $\text{NQL}(Q_{\text{HorFuncChallenge}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ .

For the converse direction, we aim at proving that there is a negative query implication on  $Q_{\text{HorFuncChallenge}}$  for those instances that encode valid tilings and are realizable. More precisely, we consider a visible instance  $\mathcal{V}$  in which the relation  $E_H$  is a function and the relation **ErrHorFun** is empty (note that the latter

condition on **ErrHorFun** is safe, in the sense that the considered instance  $\mathcal{V}$  could be obtained from a valid tiling and, being realizable, could be used to witness a negative query implication). We claim that  $\text{NQI}(Q_{\text{HorFuncChallenge}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ . Consider an arbitrary full instance  $\mathcal{F}$  that agrees with  $\mathcal{V}$  on the visible part and satisfies the constraints in  $\mathcal{C}$ , and suppose by way of contradiction that  $Q_{\text{HorFuncChallenge}}$  holds on  $\mathcal{F}$ . Then,  $\mathcal{F}$  would contain the following facts, for a triple of nodes  $x, y, y'$ : **HorFuncChallenge**( $x, y, x, y'$ ),  $E_H(x, y)$ ,  $E_V(x, y')$ . On the other hand,  $\mathcal{F}$  cannot contain the fact **HorFuncChallenge**( $x, y, x', y$ ), as otherwise this would imply the presence of the visible fact **ErrHorFun**( $x, y, x'$ ). From this we conclude that  $y \neq y'$ , which contradicts the functionality of  $E_H$ .

Very similar constructions and arguments can be used to enforce single successors in  $E_V$ , single predecessors in  $E_H$  and  $E_V$ , as well as confluence of  $E_H$  and  $E_V$ .

We now explain how we enforce the existential properties of the grid, such as  $E_H$  being non-empty. We introduce two binary relations **HorEmptyError** and **HorEmptyHiddenError**, where the former is visible and the latter is hidden, and constraint them via the CQ-view definition

$$\text{HorEmptyError} \leftrightarrow \exists x y ( E_H(x, y) \wedge \text{HorEmptyHiddenError} ) .$$

We add as a conjunct of our query the following UCQ:

$$Q_{\text{HorEmptyError}} = \text{HorEmptyError} \vee \text{HorEmptyHiddenError} .$$

Below, we show how this enforces non-emptiness of  $E_H$ .

Suppose that  $\mathcal{V}$  is an  $\mathbf{S}_v$ -instance such that  $\text{NQI}(Q_{\text{HorEmptyError}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ . We show that in this case the relation  $E_H$  is non-empty. First, note that the fact **HorEmptyError** must not appear in  $\mathcal{V}$ , since otherwise all full instances extending  $\mathcal{V}$  would satisfy  $Q_{\text{HorEmptyError}}$  (as  $\mathcal{V}$  is realizable, there is at least one such full instance). If  $E_H$  were empty, we could set **HorEmptyHiddenError** to non-empty and thus get a contradiction of  $\text{NQI}(Q_{\text{HorEmptyError}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ .

For the converse direction, we consider a visible instance  $\mathcal{V}$  in which the relation  $E_H$  is non-empty and **HorEmptyError** is empty (again, such an instance can be obtained from a valid tiling of the infinite grid and thus can be used to witness a negative query implication). In any full instance that agrees with  $\mathcal{V}$  on the visible part, **HorEmptyHiddenError** must agree with **HorEmptyError**, and hence must be empty. This implies that the query  $Q_{\text{HorEmptyError}}$  is violated, whence  $\text{NQI}(Q_{\text{HorEmptyError}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ .

Besides requiring that  $E_H$  and  $E_V$  are non-empty, we must also guarantee that for every pair  $(x, y) \in E_H$  (resp.,  $(x, y) \in E_V$ ), there is a pair  $(y, z) \in E_V$  (resp.,  $(y, z) \in E_H$ ). Note that once we have performed this, functionality and confluence will ensure that  $E_H$  and  $E_V$  correctly encode the horizontal and vertical edges of the grid. We explain how to enforce that every pair  $(x, y) \in E_H$  has a successor pair  $(y, z) \in E_V$  – a similar construction can be given for the symmetric property. We add to our schema another visible relation **HorSuccError**

of arity 0, and a hidden relation **HorSuccHiddenError** of arity 1. The associated CQ-view definition is

$$\text{HorSuccError} \rightarrow \exists x y z E_H(x, y) \wedge \text{HorSuccHiddenError}(y) \wedge E_V(y, z) .$$

Moreover, we add as a conjunct of our query the following UCQ:

$$Q_{\text{HorSuccError}} = \text{HorSuccError} \vee \left( \exists x y E_H(x, y) \wedge \text{HorSuccHiddenError}(y) \right) .$$

We show how this enforces the desired property.

Suppose that there is a visible instance  $\mathcal{V}$  such that  $\text{NQL}(Q_{\text{HorSuccError}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ . First, observe that the visible relation **HorSuccError** must be empty, as otherwise all extensions of  $\mathcal{V}$  would satisfy  $Q_{\text{HorSuccError}}$ . Now, suppose, by way of contradiction, that there is a pair  $(x, y) \in E_H$  that has no successor pair  $(y, z) \in E_V$ . In this case, we can construct a full instance that extends  $\mathcal{V}$  with the hidden fact **HorLabelHiddenError**( $y$ ). This full instance has  $\mathcal{V}$  as visible part and satisfies the constraints and the query  $Q_{\text{HorSuccError}}$ . As this contradicts the hypothesis  $\text{NQL}(Q_{\text{HorSuccError}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ , we conclude that for every pair  $(x, y) \in E_H$ , there is a successor pair  $(y, z) \in E_V$ .

Conversely, consider a visible instance  $\mathcal{V}$  that represents a correct encoding of the infinite grid and where the visible relation **HorSuccError** is empty. In any full instance that agrees with  $\mathcal{V}$  on the visible part, **HorSuccError** must be the same as  $\exists x y z E_H(x, y) \wedge \text{HorSuccHiddenError}(y) \wedge E_V(y, z)$ . In particular, because every node has both a successor in  $E_H$  and a successor in  $E_V$ , this implies that the hidden relation **HorSuccHiddenError** cannot contain the node  $y$ , for any pair  $(x, y) \in E_H$ . Hence the query  $Q_{\text{HorSuccError}}$  is necessarily violated, and this proves that  $\text{NQL}(Q_{\text{HorSuccError}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ .

Now that enforced a grid-like structure on the relations  $E_H$  and  $E_V$ , we consider the relations  $U_t$  that encode a candidate tiling function. Using similar techniques, we can ensure that every node of the grid has an associated tile. More precisely, we enforce that, for every pair  $(x, y) \in E_H$ , the element  $x$  must appear in also appears in  $U_t$ , for some tile  $t \in T$ . We add a visible relation **HorLabelError** $_t$  of arity 0 for each tile  $t \in T$  and a hidden relation **HorLabelHiddenError** of arity 1. The associated CQ-view definitions are of the form

$$\text{HorLabelError}_t \leftrightarrow \exists x y E_H(x, y) \wedge \text{HorLabelHiddenError}(x) \wedge U_t(x) .$$

We add as conjunct of our query the following UCQ:

$$Q_{\text{HorLabelError}} = \bigvee_{t \in T} \exists x y (\text{HorLabelError}_t(x, y)) \vee (E_H(x, y) \wedge \text{HorLabelHiddenError}(x)) .$$

We prove that the above definitions enforce that all nodes that appearing on the first column of the relation  $E_H$  have at least one associated tile.

Consider a visible instance  $\mathcal{V}$  such that  $\text{NQL}(Q_{\text{HorLabelError}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ . For each tile  $t$ , the visible relation **HorLabelError** $_t$  must be empty, as otherwise all extensions of  $\mathcal{V}$  would satisfy  $Q_{\text{HorLabelError}}$ . Suppose, by way of contradiction, that there is a node  $x$  that appears on the first column of the visible relation



$E_H$ , but does not appear in any relation  $U_t$ , with  $t \in T$ . We can construct a full instance where the relation `HorLabelHiddenError` contains the element  $x$ . This instance would then satisfy the query  $Q_{\text{HorLabelError}}$ , thus contradicting  $\text{NQI}(Q_{\text{HorLabelError}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ .

For the converse, consider a visible instance  $\mathcal{V}$  in which the relation  $E_H$  is non-empty (as enforced in the previous steps) and, for all pairs  $(x, y) \in E_H$ , there is a tile  $t \in T$  such that  $x \in U_t$ . Furthermore, assume that all the relations `HorLabelErrort`, with  $t \in T$ , in this visible instance are empty. Note that such an instance  $\mathcal{V}$  is realizable and hence can be obtained from a valid tiling (if there is any) and used as a witness of a negative query implication. In every full instance that agrees with  $\mathcal{V}$  and satisfies the constraints, `HorLabelErrort` must be the same as  $\exists x y \text{ HorLabelHiddenError}(x) \wedge E_H(x, y) \wedge U_t(x)$ . In particular, because every node is associated with some tile, this implies that the hidden relation `HorLabelHiddenError` cannot contain the node  $x$ , for any pair  $(x, y) \in E_H$ . Hence the query  $Q_{\text{HorLabelError}}$  is necessarily violated, and this proves that  $\text{NQI}(Q_{\text{HorLabelError}}, \mathcal{C}, \mathbf{S}, \mathcal{V}) = \text{true}$ .

We also need to guarantee that each node has at most one associated tile. This property can be easily enforced by the subquery

$$Q_{\text{TwoLabelsError}} = \bigvee_{t \neq t'} \exists x U_t(x) \wedge U_{t'}(x) .$$

Finally, we enforce that the encoded tiling function respects the horizontal and vertical constraints using the following UCQ:

$$Q_{\text{ConstraintError}} = \bigvee_{(t, t') \notin H} \left( \exists x y E_H(x, y) \wedge U_t(x) \wedge U_{t'}(y) \right) \vee \bigvee_{(t, t') \notin V} \left( \exists x y E_V(x, y) \wedge U_t(x) \wedge U_{t'}(y) \right) .$$

Summing up, if we let  $Q$  be the conjunction of all previous queries, we know that  $\exists \text{NQI}(Q, \mathcal{C}, \mathbf{S}) = \text{true}$  if and only if there exists a valid tiling of the infinite grid  $\mathbb{N} \times \mathbb{N}$ .  $\square$

## 6 Conclusions

This work gives a detailed examination of disclosure of query results from schemas with hidden relations, where disclosure arises from the presence of constraints in expressive integrity constraint languages. In future work we will look at mechanisms for “restricted access” that are finer-grained than just exposing the full contents of a subset of the schema relations, such as language-based restrictions (user can pose queries within a certain language).

## References

- [AD98] S. Abiteboul and O. Duschka. Complexity of answering queries using materialized views. In *PODS*, 1998.

- [AHV95] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.
- [BCtCB15] M. Benedikt, T. Colcombet, B. ten Cate, and M. Vanden Boom. The complexity of boundedness for guarded logics, 2015. Available at [www.cs.ox.ac.uk/michael.benedikt/papers/gnfpb.pdf](http://www.cs.ox.ac.uk/michael.benedikt/papers/gnfpb.pdf).
- [BGO10] V. Bárány, G. Gottlob, and M. Otto. Querying the guarded fragment. In *LICS*, 2010.
- [BLMS09] J.-F. Baget, M. Leclère, M.-L. Mugnier, and E. Salvat. Extending decidable cases for rules with existential variables. In *IJCAI*, 2009.
- [BtCO12] V. Bárány, B. ten Cate, and M. Otto. Queries with guarded negation. In *VLDB*, 2012.
- [BtCS11] V. Bárány, B. ten Cate, and L. Segoufin. Guarded negation. In *ICALP*, 2011.
- [CFP84] M. Casanova, R. Fagin, and C. Papadimitriou. Inclusion dependencies and their interaction with functional dependencies. *JCSS*, 28(1):29–59, 1984.
- [CGK13] A. Calì, G. Gottlob, and M. Kifer. Taming the infinite chase: Query answering under expressive relational constraints. *JAIR*, 48:115–174, 2013.
- [CK90] Chen Chung Chang and H. Jerome Keisler. *Model Theory*. North-Holland, 1990.
- [CY14] R. Chirkova and T. Yu. Obtaining information about queries behind views and dependencies. *CoRR*, abs/1403.5199, 2014.
- [FG10a] W. Fan and F. Geerts. Capturing missing tuples and missing values. In *PODS*, 2010.
- [FG10b] W. Fan and F. Geerts. Relative information completeness. *ACM TODS*, 35(4):27, 2010.
- [FIS11] E. Franconi, Y. Ibáñez-García, and I. Seylan. Query answering with DBoxes is hard. *ENTCS*, 278:71–84, 2011.
- [GM14] T. Gogacz and Jerzy Marcinkowski. All-instances termination of chase is undecidable. In *ICALP*, 2014.
- [GP03] G. Gottlob and C. Papadimitriou. On the complexity of single-rule datalog queries. *Inf. Comp.*, 183, 2003.
- [JK84] D. S. Johnson and A. C. Klug. Testing Containment of Conjunctive Queries under Functional and Inclusion Dependencies. *JCSS*, 28(1):167–189, 1984.

- [KLWW13] B. Konev, C. Lutz, D. Walther, and F. Wolter. Model-theoretic inseparability and modularity of description logic ontologies. *Artif. Intell.*, 203:66–103, 2013.
- [KUB<sup>+</sup>12] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. Query-based data pricing. In *PODS*, 2012.
- [LSW12] C. Lutz, I. Seylan, and F. Wolter. Mixing open and closed world assumption in ontology-based data access: Non-uniform data complexity. In *Description Logics*, 2012.
- [LW07] C. Lutz and F. Wolter. Conservative extensions in the lightweight description logic EL. In *CADE*, 2007.
- [MG10] B. Marnette and F. Geerts. Static analysis of schema-mappings ensuring oblivious termination. In *ICDT*, 2010.
- [MS07] G. Miklau and D. Suciu. A formal analysis of information disclosure in data exchange. *JCSS*, 73(3):507–534, 2007.
- [Var98] M. Y. Vardi. Reasoning about the past with two-way automata. In *ICALP*, 1998.
- [ZM05] Z. Zhang and A. O. Mendelzon. Authorization views and conditional query containment. In *ICDT*, 2005.